Parametrizations of Discrete Graphical Models

Robin J. Evans
`www.stat.washington.edu/∼rje42`

10th August 2011

Outline

1. Introduction
   - Graphical Models
   - Acyclic Directed Mixed Graphs
   - Two Problems
2. Ingenuous Parametrization
3. Parsimony and Model Selection
4. Variation Independence

Outline

Multivariate Statistics

Take random vectors $X_V^1, X_V^2, \ldots, X_V^n$, with components indexed by $V = \{1, \ldots, k\}$.

We wish to investigate the joint distribution of components of X_Vs

$$f(X_1, \ldots, X_k).$$

Might impose structure:

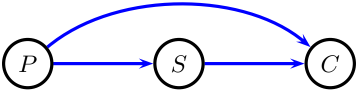
- for scientific considerations;
- for prediction (under additional covariates);
- simply to reduce the parameter count.

Graphs

Let P = cigarette price, S = smoking rate, C = lung cancer rate, with some joint distribution:

$$f(P, S, C) = f_P(P)\, f_S(S \mid P)\, f_C(C \mid \cancel{P}, S)$$

Can represent this graphically:



This is a **directed acyclic graph**.

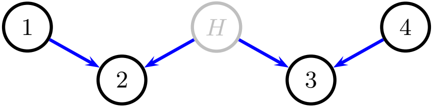
This approach creates **conditional independences**:

$$P \perp\!\!\!\perp C \mid S.$$

These can be read off the graph (global Markov property).

Latent Variables

A more complicated DAG, with a latent variable H:



This gives observable conditional independences

$$X_1 \perp\!\!\!\perp X_3, X_4 \qquad\qquad X_4 \perp\!\!\!\perp X_1, X_2. \tag{$*$}$$

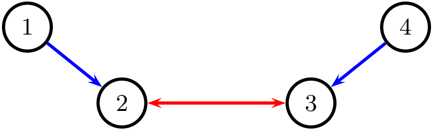
No fully observed DAG encodes precisely $(*)$.

However, latent variable models

- are not always identified;
- are not curved exponential families;
- do not have nice statistical properties.

Acyclic Directed Mixed Graphs

Replacing latents with **bidirected** edges leads to an **acyclic directed mixed graph** (ADMG).



The ADMG encodes $(*)$, making no assumptions about H.

For the class of discrete ADMGs:

- each model represents a curved exponential family;
- everything is fully identified;
- Markov property closed under marginalization.

Notation

We assume X_V are **binary**, from strictly positive distribution P.

Data in form of contingency table with 2^k cells.

Extension to general finite discrete case is easy.

Subvectors: for $A \subseteq V$, write $X_A \equiv (X_v)_{v \in A}$.

Probabilities:

$$\begin{aligned} p_{011} &\equiv P(X_1 = 0, X_2 = 1, X_3 = 1) \\ p_{0\cdot 1} &\equiv P(X_1 = 0, X_3 = 1). \end{aligned}$$

Parametrizations

To work with an ADMG model, need a **parametrization**: smooth bijective map from set of distributions in model to open parameter space $Q \subseteq \mathbb{R}^q$.

Existing parametrization of Richardson (2009) for ADMGs uses **heads** (H) conditional on **tails** (T):

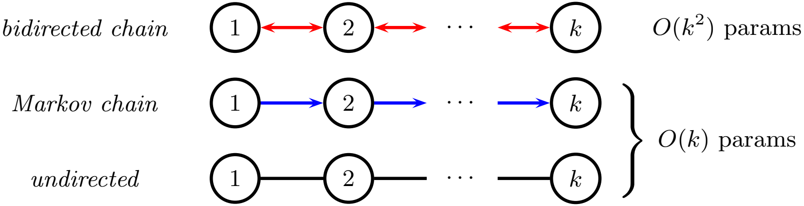
$$P(X_H = \mathbf{0} \,|\, X_T = i_T), \qquad i_T \in \mathfrak{X}_T.$$

Conditional probabilities give a smooth map, fully identifiable parameters.

Multi-linear map back to joint probabilities.

Problem 1: Parsimony

Number of parameters can be large even for sparse graphs:

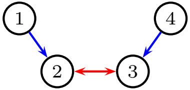


Includes high order interactions which may not be needed (**not** assuming stationarity).

No obvious way to make parsimonious sub-models using Richardson's parametrization.

Other graphical model classes have methods for parsimonious sub-models (e.g. undirected graphs → Boltzmann Machines).

Problem 2: Variation Dependence



The ten parameters for our model are

$$\overset{(1)}{P(X_1 = 0)} \quad \overset{(1)}{P(X_4 = 0)} \quad \overset{(2)}{P(X_2 = 0 \mid X_1 = x_1)} \quad \overset{(2)}{P(X_3 = 0 \mid X_4 = x_4)}$$

$$\overset{(4)}{P(X_2 = 0,\ X_3 = 0 \mid X_1 = x_1, X_4 = x_4)}\ .$$

Note the variation dependence, e.g.:

$$P(X_2 = 0,\ X_3 = 0 \mid X_1 = x_1, X_4 = x_4) \leq P(X_2 = 0 \mid X_1 = x_1).$$

Variation independence means that prior specification, parameter interpretation, regression modelling all become easier.

Alternatively, could use conditional odds-ratios:

$$\frac{P(X_2 = 0,\, X_3 = 0 \,|\, x_1,\, x_4) \cdot P(X_2 = 1,\, X_3 = 1 \,|\, x_1,\, x_4)}{P(X_2 = 1,\, X_3 = 0 \,|\, x_1,\, x_4) \cdot P(X_2 = 0,\, X_3 = 1 \,|\, x_1,\, x_4)}$$

Does this work more generally?

Fitting

Variation dependence also makes it harder to create a fitting algorithm. Nevertheless:

Theorem (Evans and Richardson, 2010)

ADMG models can be fitted (by Maximum Likelihood Estimation) using a block co-ordinate updating scheme with gradient ascent.

Outline

Marginal Log-Linear Parameters

For $L \subseteq M \subseteq V$, define

$$\lambda_L^M \equiv \frac{1}{2^{|M|}} \sum_{j_M \in \{0,1\}^{|M|}} (-1)^{|j_L|} \log P(X_M = j_M),$$

the marginal log-linear parameter for effect L in margin M. (Bersgma and Rudas, 2002).

This is just coefficient for set L in ordinary log-linear expansion for margin M. Examples:

$$\lambda_1^1 = \frac{1}{2} \log \frac{p_{0\cdot\cdot}}{p_{1\cdot\cdot}} \qquad \lambda_{123}^{123} = \frac{1}{8} \log \frac{p_{000}\, p_{110}\, p_{101}\, p_{011}}{p_{100}\, p_{010}\, p_{001}\, p_{111}}$$

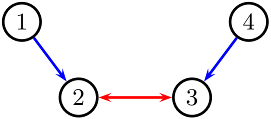
$$\lambda_1^{12} = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{01\cdot}}{p_{10\cdot}\, p_{11\cdot}} \qquad \lambda_{12}^{12} = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{11\cdot}}{p_{10\cdot}\, p_{01\cdot}}$$

The Ingenuous Parametrization

For an ADMG $\mathcal{G}$, take

$$\Lambda(\mathcal{G}) \equiv \{\lambda_A^{HT} \,|\, H \subseteq A \subseteq H \cup T,\ H \text{ a head}\}.$$

Call these the **ingenuous parameters** of $\mathcal{G}$. Example:



H	1	4	2	3	23
T	$\emptyset$	$\emptyset$	1	4	14

Richardson	**Ingenuous**
$P(X_1 = 0)$	λ_1^1
$P(X_4 = 0)$	λ_4^4
$P(X_2 = 0 \,\vert\, X_1 = x_1)$	λ_2^{12}, λ_{12}^{12}
$P(X_3 = 0 \,\vert\, X_4 = x_4)$	λ_3^{34}, λ_{34}^{34}
$P(X_2 = 0, X_3 = 0 \,\vert\, X_1 = x_1, X_4 = x_4)$	λ_{23}^{1234}, λ_{123}^{1234}, λ_{234}^{1234}, λ_{1234}^{1234}

Parametrization Result

Theorem (Evans, 2011)

The ingenuous parameters for an ADMG $\mathcal{G}$ smoothly parametrize all distributions obeying the global Markov property with respect to $\mathcal{G}$.

Outline

Problem 1: Sub-Models

The new parametrization makes it easy to produce parsimonious sub-models.

For the chain model, the ingenuous parameters are:

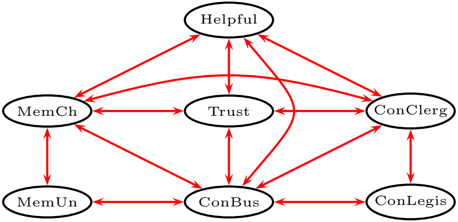
$$\lambda_1^1, \lambda_2^2, \cdots, \lambda_k^k; \quad \lambda_{12}^{12}, \lambda_{23}^{23}, \cdots, \lambda_{k-1,k}^{k-1,k}; \quad \cdots \quad \lambda_{12\cdots k}^{12\cdots k}$$

To make the model sparser (say grow linearly with the length of the chain) we could set $\lambda_M^M = 0$ for $|M| > l$, some l.

GSS Example

Drton and Richardson (2008) fit bidirected graphs to binary data from 7 questions of the General Social Survey (GSS) ($n = 13{,}486$).

They select the following model with 101 parameters (dev. 32.7 on 26 d.o.f.):



By comparison, a well fitting undirected model can be found with 39 parameters (dev. 87.6 on 88 d.o.f).

Eliminating ≥ 4-way interactions on the bidirected model gets us to 46 params (84.18 on 81 d.o.f); can do even better by removing some 3-way parameters.

Automatic Approaches

The previous approach for removing higher order interactions is *ad hoc*. Would be nice to have method for automatic parameter selection and estimation.

We desire estimator $\tilde{\boldsymbol{\theta}}$ to have **oracle** properties: as $n \to \infty$,

- $\tilde{\boldsymbol{\theta}}$ sets correct parameters to zero eventually;
- non-zero parameters are estimated (Cramér-Rao) efficiently.

Best known simultaneous method is Tibshirani's lasso: find value $\tilde{\boldsymbol{\theta}}$ which maximizes

$$l_n(\boldsymbol{\theta}) - \nu_n \sum_j |\theta_j|$$

for some $\nu_n > 0$. L_1-penalty shrinks some parameters to be exactly zero.

Unfortunately, ordinary lasso is not known to be oracle.

The Adaptive Lasso

Zou (2006) proposed the **adaptive lasso**: maximize

$$l_n(\boldsymbol{\theta}) - \nu_n \sum_j w_j |\theta_j|$$

for weights $w_j = |\hat{\theta}_j|^{-\gamma}$ and $\gamma > 0$. Here $\hat{\boldsymbol{\theta}}$ is some consistent estimator (e.g. MLE).

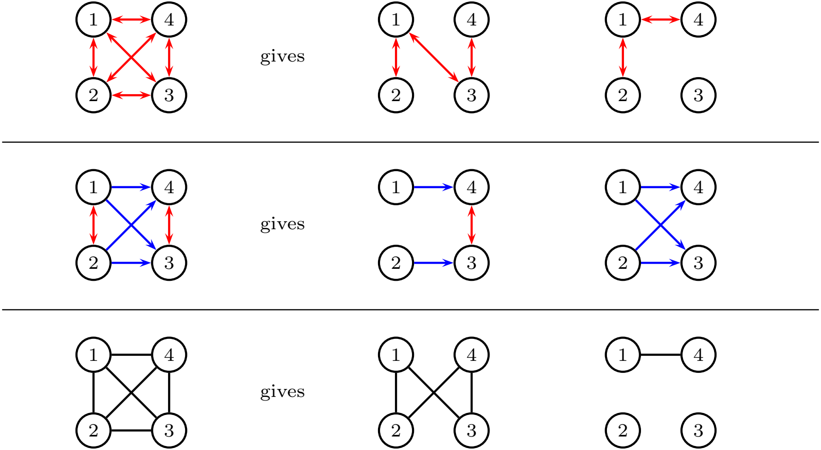
Zou shows this is oracle for linear regression models.

Theorem (Evans, 2011)

Let $\nu_n = o(\sqrt{n})$ and $\nu_n n^{\frac{\gamma-1}{2}} \to \infty$. Then the adaptive lasso estimator is oracle for marginal log-linear parametrizations.

Model Selection

Provides fully automatic method for model selection within subclass.

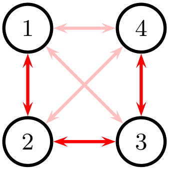


Model returned is not necessarily graphical.

Grouped lasso might be applied to select from particular subsets of models.

Simulations

Generate a probability distribution from model:



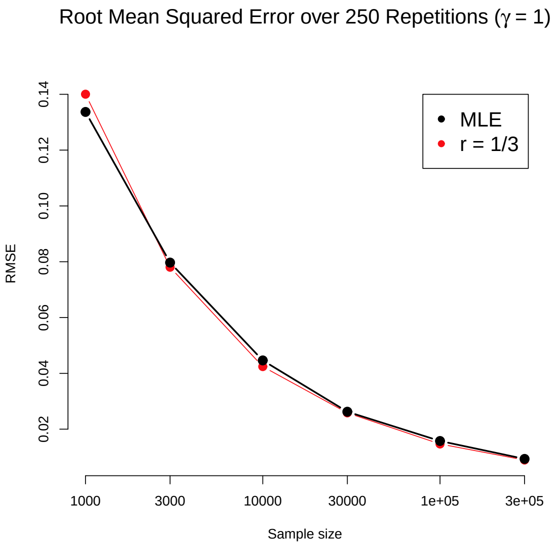
Generate data set of size n from the distribution.

Try to recover correct model and distribution using the (adaptive) lasso with cross validation.

Repeated $N = 250$ times for various $\gamma \in \{0, \frac{1}{2}, 1\}$ and penalties $\nu_n = Cn^r$ with rates $r \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$, where C chosen by cross-validation.

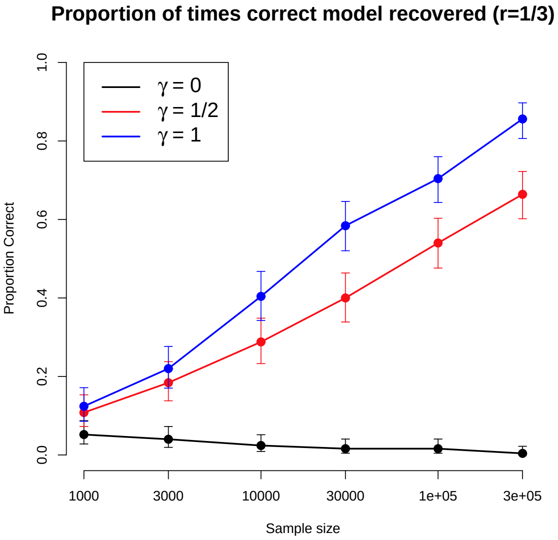
Sample sizes $n = 10^3$ up to 3×10^5.

Results



Root Mean Squared Error over 250 Repetitions (γ = 1)
RMSE
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1000
3000
10000
30000
1e+05
3e+05
Sample size
MLE
r = 1/3

Results



Proportion of times correct model recovered (r=1/3)
γ = 0
γ = 1/2
γ = 1
Proportion Correct
0.0
0.2
0.4
0.6
0.8
1.0
1000
3000
10000
30000
1e+05
3e+05
Sample size

Outline

Problem 2: Variation Independence

We can characterize when the ingenuous parametrization is variation independent.

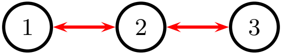
Theorem (Evans, 2011)

Ingenuous parametrization for an ADMG $\mathcal{G}$ is variation independent iff $\mathcal{G}$ has no heads of size ≥ 3.

Note that this includes all DAGs and, for example,

Variation Dependence

What goes wrong with heads of size 3?



Ingenuous parameters are

$$\lambda_1^1, \quad \lambda_2^2, \quad \lambda_{12}^{12}, \quad \lambda_3^3, \quad \lambda_{23}^{23}, \quad \lambda_{123}^{123}.$$

Need to sequentially choose parameter values.

Working marginally, could make $\mathrm{Corr}(X_1, X_2)$ and $\mathrm{Corr}(X_2, X_3)$ very large using λ_{12}^{12} and λ_{23}^{23}. If these correlations are too high, X_1 and X_3 cannot be independent.

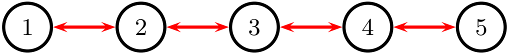
Need to ensure Fréchet bounds are not violated (Dobra and Feinberg, 2000).

For VI parametrization replace λ_{23}^{23} with non-marginal parameter

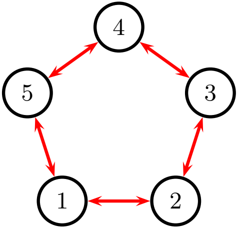
$$\tilde{\lambda}_{23}^{23} = \lambda_{23}^{23} + \frac{1}{4} \log \frac{(p_{000} + p_{111})(p_{011} + p_{100})}{(p_{010} + p_{101})(p_{001} + p_{110})}.$$

Variation Dependence

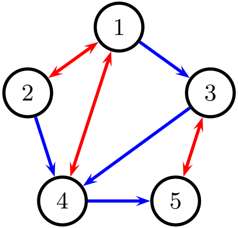
This approach gives a variation independent parametrization for the bidirected 5-chain (and 6-chain).



However it doesn't work for all models:



Variation Independence in General

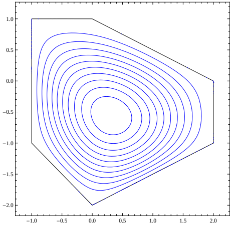


Pick a topological ordering of vertices, $1, 2, \ldots, k$.

By induction assume model over $1, \ldots, k-1$ has VI parametrization.

Conditional on parameters for first $k-1$ vertices, Richardson's parameters for k are linearly constrained.

Valid range is then a (non-empty) convex polytope, which can be mapped onto a ball.



Variation Independence in General

Using this approach:

Theorem (Evans, 2011)

We can construct variation independent parametrizations for all ADMG models.

Can also ensure that setting parameters to zero has some meaning (e.g. context specific CI).

Summary

We have:

- presented Richardson's parametrization for discrete acyclic directed mixed graphs;
- given a new parametrization based on marginal log-linear parameters;
- shown how this parametrization may be used to create parsimonious sub-models, and used for automatic model selection;
- characterized variation independence for the new parametrization;
- shown that all ADMG models have a variation independent parametrization.

Acknowledgements

Thomas Richardson.

Committee members: Adrian Dobra, Brian Flaherty, Peter Hoff, Steffen Lauritzen and James Robins.

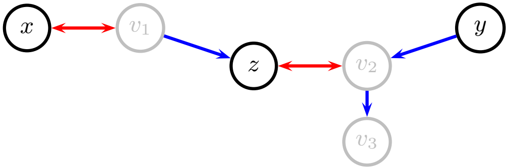
Thanks also to Antonio Forcina and Tamás Rudas for discussions, helpful comments and code.

Thank you!

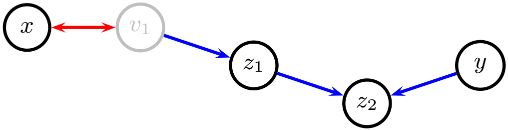
m-separation

Two vertices x and y are m-separated by a set Z if all paths from x to y are blocked by Z.

Either: at least one collider is not conditioned upon, and nor are any of its descendants:



Or: at least one non-collider is conditioned upon:



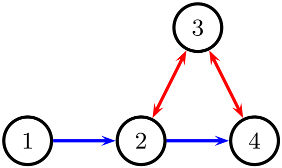
m-separation extends to sets X and Y if every $x \in X$ and $y \in Y$ are m-separated.

Global Markov Property

Let P be a distribution over the vertices of $\mathcal{G}$. The global Markov property (GMP) for ADMGs states that

$$X \text{ m-separated from } Y \text{ by } Z \quad \Longrightarrow \quad X \perp\!\!\!\perp Y \mid Z \; [P]$$

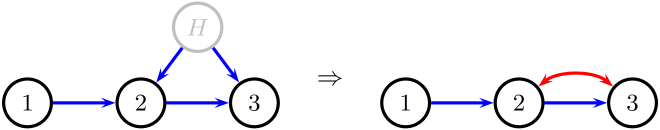
Example:



Here $1 \perp\!\!\!\perp 4 \mid 2$ and $1 \perp\!\!\!\perp 3$.

Markov Property Closure

Global Markov property for ADMGs is closed under marginalization (preserves conditional independences):



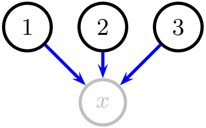
However in the DAG,

$$P(X_2 = 0,\, X_3 = 0 \,|\, X_1 = 0) + P(X_2 = 0,\, X_3 = 1 \,|\, X_1 = 1) \leq 1$$

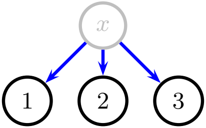
(and 3 other inequalities).

Definitions

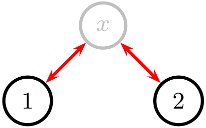
parents $\mathrm{pa}_{\mathcal{G}}(x)$



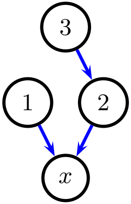
children $\mathrm{ch}_{\mathcal{G}}(x)$



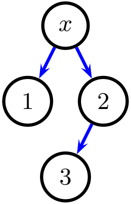
spouses $\mathrm{sp}_{\mathcal{G}}(x)$



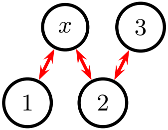
ancestors $\mathrm{an}_{\mathcal{G}}(x)$



descendants $\mathrm{de}_{\mathcal{G}}(x)$



district $\mathrm{dis}_{\mathcal{G}}(x)$



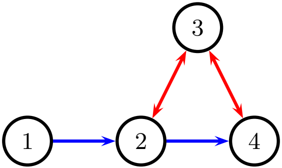
Parametrizing ADMGs

Richardson (2009) gives a parametrization of discrete distributions obeying the global Markov property for an ADMG.

Define a *head* H to be any set of vertices which is

(i) connected by $\leftrightarrow$-arrows in $\mathrm{an}_{\mathcal{G}}(H)$;

(ii) *barren*: no element of H is an ancestor of any other.



H	1	2	3	23	4	34
T	$\emptyset$	1	$\emptyset$	1	2	12

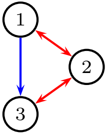
For a head H, the corresponding *tail* is the set of ancestors which are connected to H by paths of colliders in $\mathrm{an}_{\mathcal{G}}(H)$. The tail is the Markov blanket for H in $\mathrm{an}_{\mathcal{G}}(H)$.

Generalized Möbius Parameters

For head-tail pair (H, T) and $i_T \in \{0,1\}^{|T|}$, let

$$q_{H|T}^{(i_T)} \equiv P(X_H = \mathbf{0} \mid X_T = i_T),$$

a **generalized Möbius parameter**. The collection of all generalized Möbius parameters is the Richardson (2009) parametrization of the ADMG.



H	1	2	12	3	23
T	$\emptyset$	$\emptyset$	$\emptyset$	1	1

$$q_1 \quad q_2 \quad q_{12}$$
$$q_{3|1}^{(0)} \quad q_{3|1}^{(1)}$$
$$q_{23|1}^{(0)} \quad q_{23|1}^{(1)}$$

Parametrization uses Möbius like expansions, e.g.

$$p_{101} = q_2 - q_{12} - q_1 \cdot q_{3|1}^{(1)} + q_1 \cdot q_{23|1}^{(1)}.$$

Parameters are variation dependent, making fitting tricky.

Fitting Algorithm

Choose a vertex v, fix parameters associated with a head not containing v (example for $v = 1$):

$$p_{101} = q_2 - q_{12} - q_1 \cdot q_{3|1}^{(1)} + q_1 \cdot q_{23|1}^{(1)}.$$

Now $\boldsymbol{p}$ is linear in remaining parameters θ^v. Constraints amount to $\boldsymbol{p} \geq 0$, so can write as $A^v \theta^v - \boldsymbol{b}^v \geq \mathbf{0}$.

Algorithm. Cycle through each vertex $v \in V$:
Step 1. Construct the constraint matrix A^v.
Step 2. Solve the non-linear program

$$\begin{array}{ll} \text{maximize} & l(\theta^v) = \sum_{\boldsymbol{i}} n_{\boldsymbol{i}} \log p_{\boldsymbol{i}}^v(\theta^v) \\ \text{subject to} & A^v \theta^v \geq \boldsymbol{b}^v. \end{array}$$

Stop when a complete cycle of V results in a sufficiently small increase in the likelihood. □

Model Classes