Marginal log-linear parameters, graphical models and model selection.

Robin J. Evans
www.statslab.cam.ac.uk/∼rje42

20th January 2012

Acknowledgements

Antonio Forcina (Perugia)

Thomas Richardson (Washington)

Outline

Model Selection

Suppose we have n i.i.d. observations of binary random vector $(X_1, X_2, \ldots, X_k)$ from unknown $P > 0$.

Want to estimate P and choose some appropriate sub-model (e.g. best undirected graphical model).

Could use e.g. BIC:

- fit each sub-model to data;
- pick model which minimises BIC.

This is **model consistent**, i.e. for sufficiently large sample size it will find smallest sub-model.

However, requires fitting each sub-model separately.
Local search would be quicker, but not guaranteed to find best model.

Can we find an estimator which does this automatically?

Parameter Estimation

Typical statistical problem: have n i.i.d. observations from $P_{\boldsymbol{\theta}}$. Wish to infer $\boldsymbol{\theta} \in \Theta$.

$$\boldsymbol{\theta} = (\ 0.1, \quad -0.3, \quad 0, \quad 0.1, \quad 0, \quad -0.9)^T$$

Maximum likelihood estimator is consistent and asymptotically normal with Cramér-Rao variance.

$$\hat{\boldsymbol{\theta}}^n = (0.13, \ -0.24, \ 0.03, \ 0.02, \ -0.04, \ -0.97)^T$$

However, $\boldsymbol{\theta}$ lies in interesting sub-model $\theta_3 = \theta_5 = 0$.

Could perform hypothesis test (e.g. likelihood ratio), fail to reject $\theta_3 = \theta_5 = 0$, and then take constrained estimate:

$$\tilde{\boldsymbol{\theta}}^n = (0.12, \ -0.22, \ 0, \ 0.03, \ 0, \ -0.94)^T$$

This is just an estimation procedure.

Model Consistency

More generally, define our model:

$$A = \{j \mid \theta_j \neq 0\} \qquad \text{true model}$$
$$\tilde{A}^n = \{j \mid \tilde{\theta}^n_j \neq 0\} \qquad \text{estimated model.}$$

Aim to:

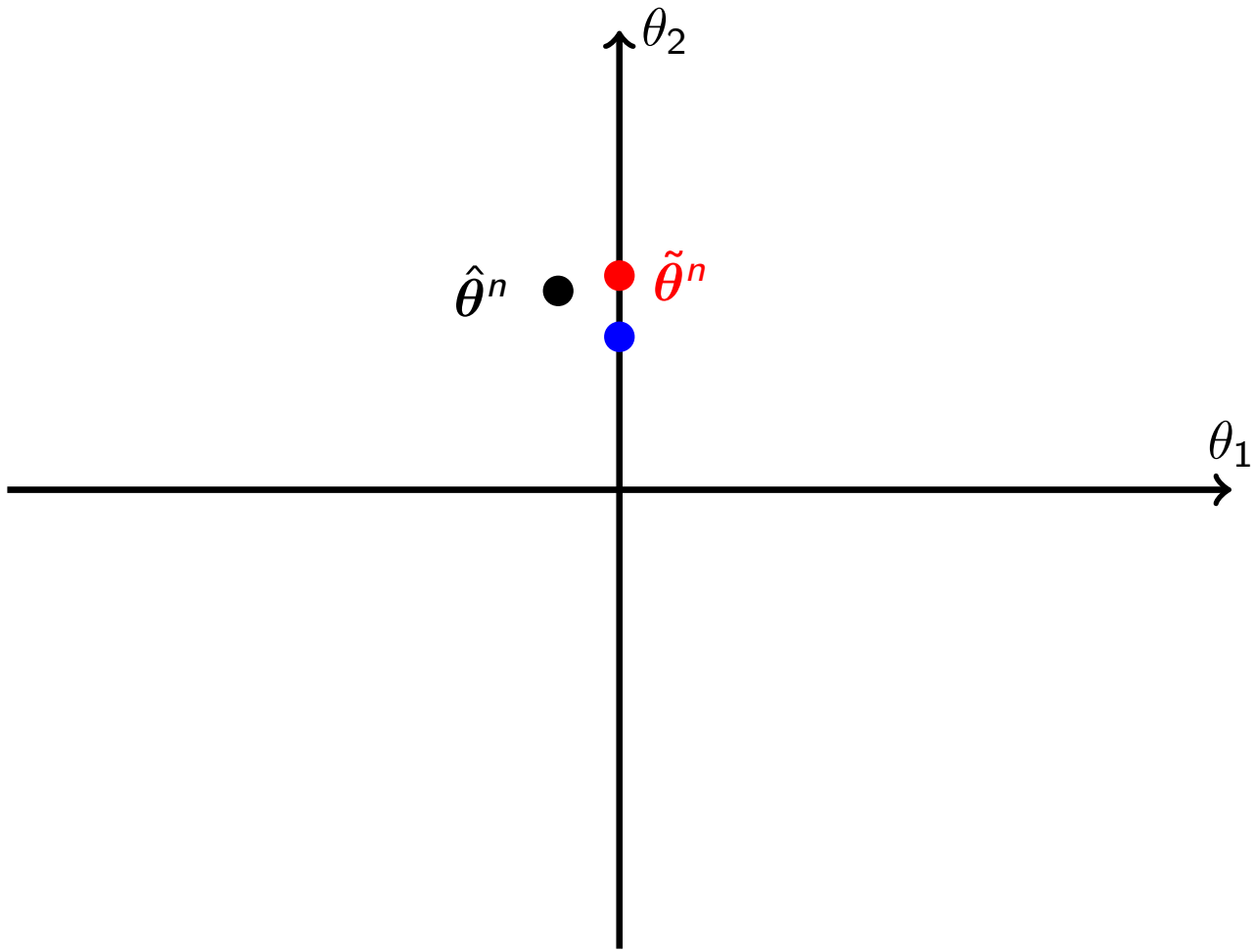
- estimate the model, A; and
- estimate the parameter vector, $\boldsymbol{\theta}$.

An estimator is $\tilde{\boldsymbol{\theta}}^n$ is **model consistent** if

$$P(\tilde{A}^n = A) \to 1.$$

Model Selection

Maximum likelihood estimators are not model consistent.



Oracle Estimators

MLE satisfies:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, I(\boldsymbol{\theta})^{-1}).$$

Our ideal $\tilde{\boldsymbol{\theta}}^n$ estimator will be **oracle**:

- model consistent: $P(\tilde{A}^n = A) \to 1$; and
- non-zero parameters $\boldsymbol{\theta}_A$ estimated efficiently:

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^n_A - \boldsymbol{\theta}_A) \xrightarrow{\mathcal{D}} N(0, I(\boldsymbol{\theta})^{AA}).$$

BIC with maximum likelihood for example, is an oracle estimator.

Want an 'automatic' oracle estimator (i.e. don't have to fit all the sub-models.

The Lasso

Ordinary lasso (Tibshirani, 1996) minimises

$$-l(\boldsymbol{\theta}) + \nu_n \sum_{j=1}^{p} |\theta_j|.$$

L_1-penalty can force some parameters to be precisely zero.

Unfortunately, the **lasso estimate is not oracle** (Zou, 2006).

The Adaptive Lasso

Zou (2006) introduced the **adaptive lasso estimator** $\tilde{\boldsymbol{\theta}}^n$, which minimises

$$-l(\boldsymbol{\theta}) + \nu_n \sum_{j=1}^{p} \hat{w}_j |\theta_j|,$$

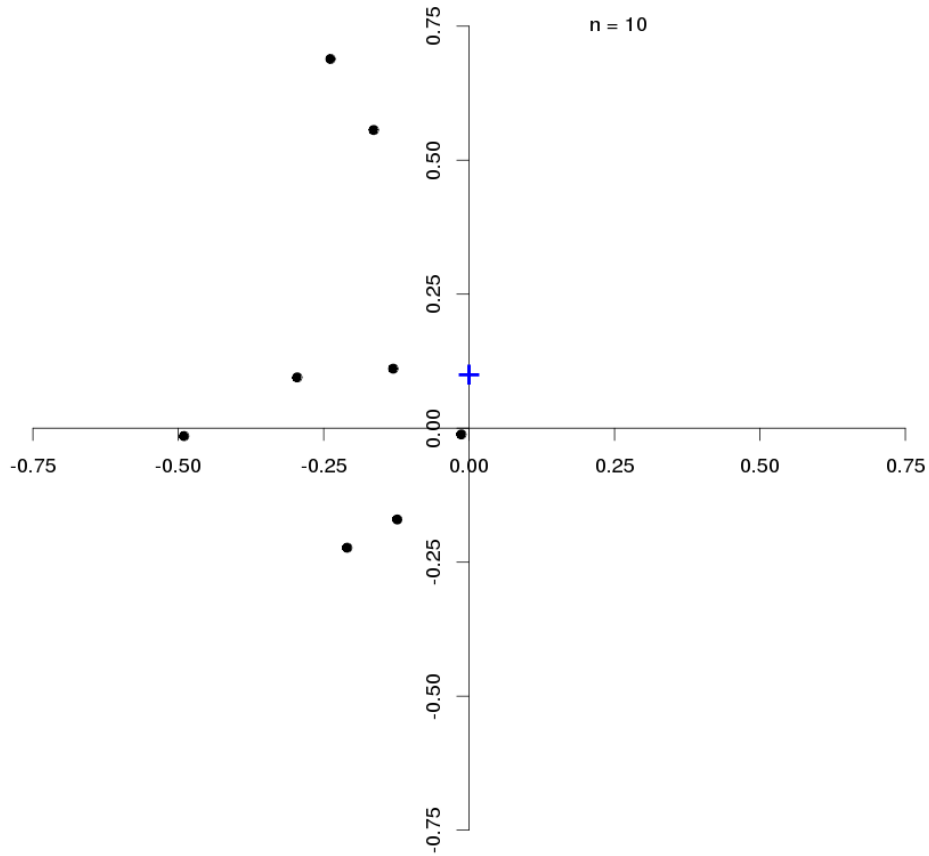
where $\hat{w}_j = |\hat{\theta}_j|^{-\gamma}$ for some consistent estimate $\hat{\theta}_j$.

The idea is that

- if $\theta_j = 0$, then $\hat{w}_j \to \infty$, and penalty diverges;
- but if $\theta_j \neq 0$, then $\hat{w}_j \to |\theta_j|^{-\gamma} < \infty$, and likelihood dominates.

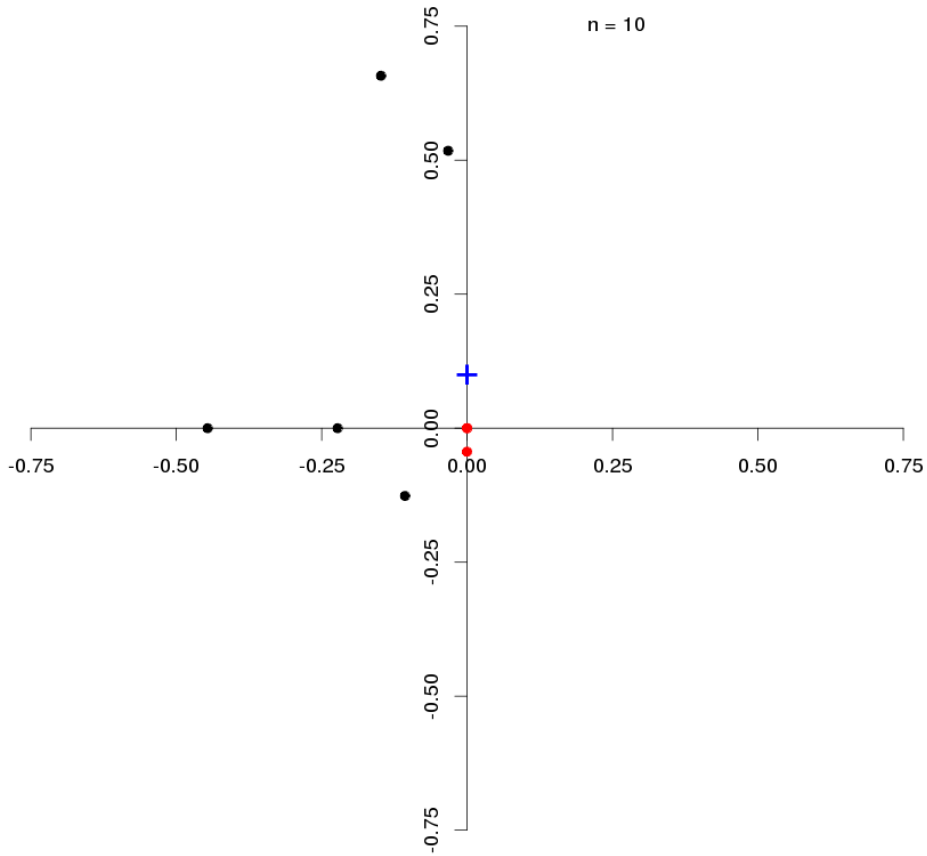
Zou shows that **adaptive lasso is oracle** for Gaussian linear regression.

Maximum Likelihood Estimate



(click to start—works in Adobe reader)

Adaptive Lasso Estimate



(click to start—works in Adobe reader)

Model Selection

Theorem (Evans, 2011)

Let $\{P_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\}$ be a parametric family obeying regularity conditions 1–5.
For suitable $\nu_n \to \infty$ and $\gamma > 0$, the adaptive lasso estimator $\tilde{\boldsymbol{\theta}}^n$ is an oracle estimator for $\boldsymbol{\theta}$:

$$P(\tilde{A}^n = A) \to 1,$$

and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^n_A - \boldsymbol{\theta}_A) \xrightarrow{\mathcal{D}} N(0, I(\boldsymbol{\theta})^{AA}).$$

Hence we can use this for discrete graphical model selection, if we can find a suitable parametrisation.

The 'Oracle' Property

The adaptive lasso is called 'oracle' because we get same asymptotic covariance as if we knew correct model in advance.

This is misleading:

- convergence is not uniform over parameter space;
- standard errors based on post model-selection estimator may be completely wrong.

However, **any** model consistent procedure has these deficiencies (Leeb and Pötscher, 2008).

We can always be 'not far away' (i.e. $o(n^{-1/2})$) from a sub-model and not realise.

Contingency Tables

Suppose we have positive joint distribution on binary (X_1, X_2, X_3).

$X_3 = 0$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	p_{000}	p_{010}
$X_1 = 1$	p_{100}	p_{110}

$X_3 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	p_{001}	p_{011}
$X_1 = 1$	p_{101}	p_{111}

Probability vector $\mathbf{p} = (p_{000}, \ldots, p_{111})$ lives in 7-dimensional probability simplex

$$\Delta_7 = \left\{ \mathbf{p} \;\middle|\; p_{ijk} > 0, \sum_{i,j,k} p_{ijk} = 1 \right\}.$$

Call this the **saturated model**.

A parametrisation is a diffeomorphism from open subset of $\mathbb{R}^7$ to Δ_7:

$$p_{000}, \quad p_{100}, \quad p_{010}, \quad p_{110}, \quad p_{001}, \quad p_{101}, \quad p_{011}.$$

Log-Linear Parameters

Can parametrise joint distribution of (X_1, X_2, X_3) with log-linear parameters rather than probabilities:

$$\log p_{x_1 x_2 x_3} = \sum_{L \subseteq \{1,2,3\}} (-1)^{|x_L|} \eta_L^{123}$$

So e.g.

$$\log p_{000} = \eta_\emptyset^{123} + \eta_1^{123} + \eta_2^{123} + \eta_{12}^{123} + \eta_3^{123} + \eta_{13}^{123} + \eta_{23}^{123} + \eta_{123}^{123}$$
$$\log p_{011} = \eta_\emptyset^{123} + \eta_1^{123} - \eta_2^{123} - \eta_{12}^{123} - \eta_3^{123} - \eta_{13}^{123} + \eta_{23}^{123} + \eta_{123}^{123}$$

Note $\eta_\emptyset^{123}$ is redundant because $\sum_{i,j,k} p_{ijk} = 1$.

This gives parameters such as

$$\eta_{23}^{123} = \frac{1}{8} \log \frac{p_{000}\, p_{100}\, p_{011}\, p_{111}}{p_{010}\, p_{110}\, p_{001}\, p_{101}}.$$

Log-Linear Parameters

The point being:

$$\eta_1^{123}, \quad \eta_2^{123}, \quad \eta_{12}^{123}, \quad \eta_3^{123}, \quad \eta_{13}^{123}, \quad \eta_{23}^{123}, \quad \eta_{123}^{123},$$

gives smooth parametrisation of the joint distribution of (X_1, X_2, X_3).

And setting any combination of these to 0 constitutes a valid non-empty sub-model.

Parametrisations

It may not be logical to treat the variables symmetrically.

Example. Consider a simple medical study:

X_1	assigned treatment	$1 =$ 'treatment', $0 =$ 'control',
X_2	took treatment	$1 =$ 'took', $0 =$ 'didn't take',
X_3	survival at 5 yrs	$1 =$ 'survived', $0 =$ 'died'.

In this case, do we really care about p_{000}?

Could use conditional probabilities instead:

$$P(X_1 = 0), \qquad P(X_2 = 0 \,|\, X_1 = x_1), \qquad P(X_3 = 0 \,|\, X_1 = x_1, X_2 = x_2).$$

When choosing a parametrisation, we might consider

- interpretation of the parameters;
- causal ordering;
- variation dependence.

Marginal Log-Linear Parameters

We can repeat the log-linear expansion in margins too:

$$\eta_1^1 = \frac{1}{2} \log \frac{p_{0++}}{p_{1++}},$$

$$\eta_2^{12} = \frac{1}{4} \log \frac{p_{00+}\, p_{10+}}{p_{01+}\, p_{11+}}, \qquad \eta_{12}^{12} = \frac{1}{4} \log \frac{p_{00+}\, p_{11+}}{p_{10+}\, p_{01+}}.$$

These are **marginal log-linear parameters**.

As with probabilities, we can mix margins to get a different (smooth) parametrisation (Bergsma and Rudas, 2002):

$$\eta_1^1, \quad \eta_2^{12}, \quad \eta_{12}^{12}, \quad \eta_3^{123}, \quad \eta_{13}^{123}, \quad \eta_{23}^{123}, \quad \eta_{123}^{123}.$$

Then from the previous example,

- no effect of assignment on treatment corresponds to $\eta_{12}^{12} = 0$;
- no effect of assignment on outcome corresponds to $\eta_{13}^{123} = \eta_{123}^{123} = 0$.

Marginal Log-Linear Parametrisations

In fact, any non-decreasing sequence of margins gives a (smooth) parametrisation of the set of positive probability distributions.

1. Take a non-decreasing sequence of margins $M_1, \ldots, M_k$, where $M_j \not\subseteq M_i$ for $i < j$, and where M_k is full margin.
   e.g. $M_1 = \{1, 2\}$, $M_2 = \{3\}$, $M_3 = \{1, 2, 3\}$.
2. Fill in remaining non-empty subsets:

M	L
$\{1, 2\}$	$\{1\}$, $\{2\}$, $\{1, 2\}$
$\{3\}$	$\{3\}$
$\{1, 2, 3\}$	$\{1, 3\}$, $\{2, 3\}$, $\{1, 2, 3\}$.

3. Take η_L^M for each (L, M)-pair.

$$\eta_1^{12}, \quad \eta_2^{12}, \quad \eta_{12}^{12}, \qquad \eta_3^3, \qquad \eta_{13}^{123}, \quad \eta_{23}^{123}, \quad \eta_{123}^{123}.$$

$M_1 = \{1, 2, 3\}$ gives ordinary log-linear parameters.

Conditional Independence

Why use marginal log-linear parameters?

- setting any sub-vector of MLL parametrisation to 0 gives a valid (non-empty) sub-model;
- some of these sub-models are meaningful:

$$X_1 \perp\!\!\!\perp X_2 \quad \Longleftrightarrow \quad \eta_{12}^{12} = 0$$

$$X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \quad \Longleftrightarrow \quad \eta_{13}^{123} = \eta_{123}^{123} = 0.$$

More generally:

Lemma (Rudas et al 2010, Forcina et al 2010)

Let P be a probability distribution parametrised by MLLPs $\boldsymbol{\eta}$. Then $X_A \perp\!\!\!\perp X_B \,|\, X_C \,[P]$ if and only if

$$\eta_{A'B'C'}^{ABC} = 0 \qquad \text{for all} \quad \begin{array}{r} \emptyset \neq A' \subseteq A, \\ \emptyset \neq B' \subseteq B, \\ C' \subseteq C. \end{array}$$

Specifying Models

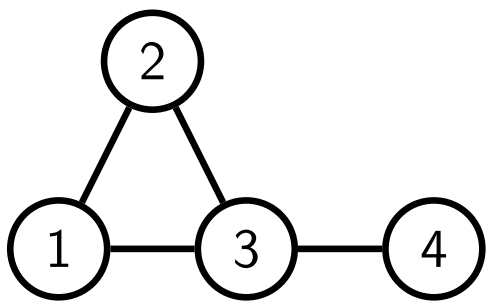
Choice of parametrisation restricts available model classes; e.g.:

$$\eta_1^{123}, \quad \eta_2^{123}, \quad \eta_{12}^{123}, \quad \eta_3^{123}, \quad \eta_{13}^{123}, \quad \eta_{23}^{123}, \quad \eta_{123}^{123}$$

- can easily specify $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$ by $\eta_{12}^{123} = \eta_{123}^{123} = 0$.
- can't easily specify $X_1 \perp\!\!\!\perp X_2$.

Log-linear models include models defined by global Markov property for undirected graph.

Set $\eta_C^V = 0$ for each set C which is not complete.

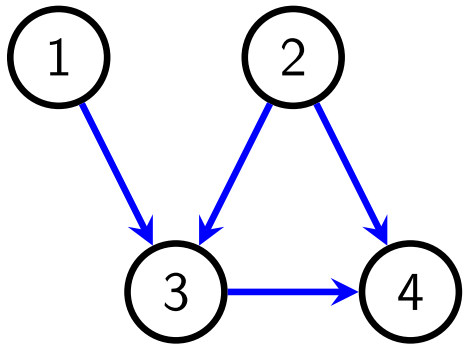


$$\eta_{14}^{1234} = \eta_{24}^{1234} = \eta_{124}^{1234} = 0$$
$$\eta_{134}^{1234} = \eta_{234}^{1234} = \eta_{1234}^{1234} = 0.$$

Model: $X_1, X_2 \perp\!\!\!\perp X_4 \,|\, X_3$ (compare with Lemma).

Directed Acyclic Graphs

Given its parents, each vertex is independent of other predecessors.



$X_2 \perp\!\!\!\perp X_1$

$X_4 \perp\!\!\!\perp X_1 \,|\, X_2, X_3$

M	L
$\{1\}$	$\{1\}$
$\{1,2\}$	$\{2\}$, $\{1,2\}$
$\{1,2,3\}$	$\{3\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$
$\{1,2,3,4\}$	$\{4\}$, $\{1,4\}$, $\{2,4\}$, $\{1,2,4\}$, …, $\{2,3,4\}$, $\{1,2,3,4\}$.

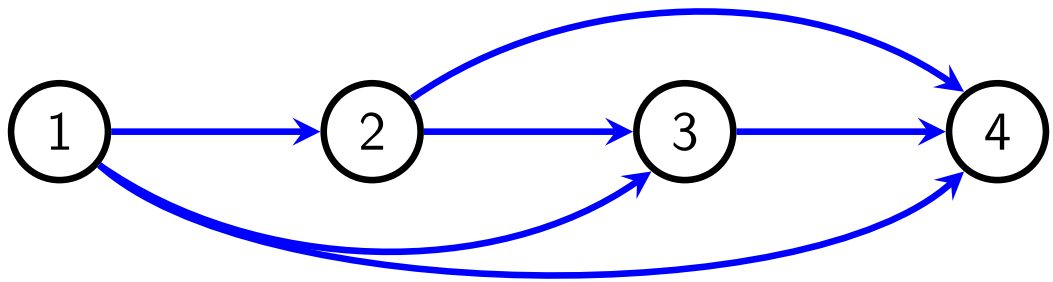
Applying the Lemma: need $\eta_{12}^{12} = \eta_{14}^{1234} = \eta_{124}^{1234} = \eta_{134}^{1234} = \eta_{1234}^{1234} = 0$.

Directed Acyclic Graphs

Thus if we use MLL parametrisation $\{1\}, \{1,2\}, \{1,2,3\}, \ldots,$

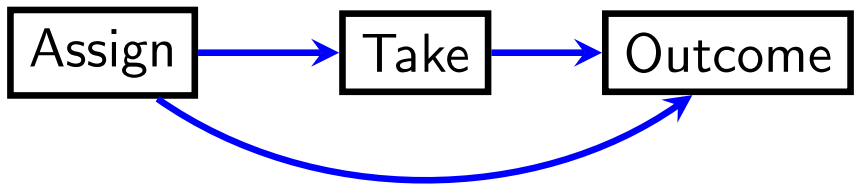
$$\eta_1^1, \quad \eta_2^{12}, \quad \eta_{12}^{12}, \quad \eta_3^{123}, \quad \ldots \quad \eta_{1234}^{1234},$$

then any DAG with this topological ordering is a linear sub-model.

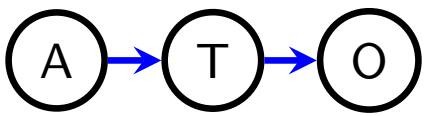


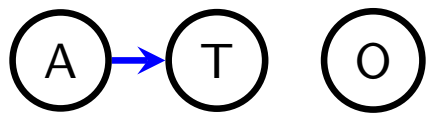
Example

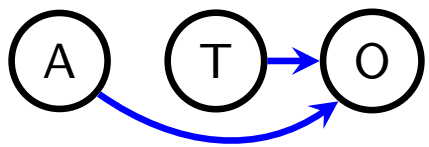
Recall our medical trial example.



Interesting sub-models correspond to sub-graphs:



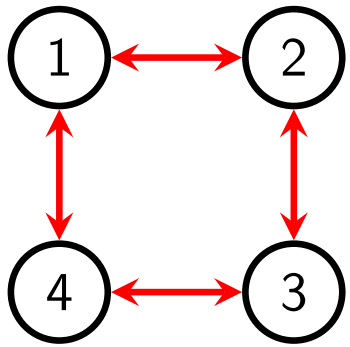




Using suggested parametrisation allows these models to be selected.

Bidirected Graphs

Vertices are marginally independent if not directly joined:



$$X_1 \perp\!\!\!\perp X_3$$

$$X_2 \perp\!\!\!\perp X_4.$$

Take **all** non-empty subsets:

M	L
$\{1\}$	$\{1\}$
$\{2\}$	$\{2\}$
$\{1,2\}$	$\{1,2\}$
$\vdots$	$\vdots$
$\{1,2,3,4\}$	$\{1,2,3,4\}$.

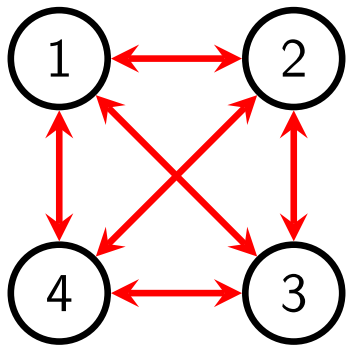
Set $\eta_D^D = 0$ if D is disconnected subset in graph ($\eta_{13}^{13} = \eta_{24}^{24} = 0$).

Bidirected Graphs

If we choose the MLL parametrisation based on all margins:

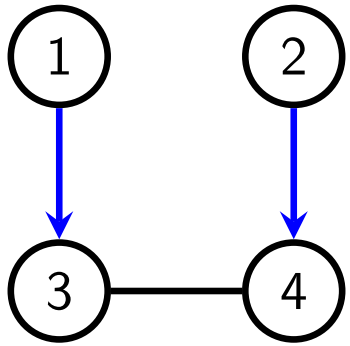
$$\eta^{1}_{1}, \quad \eta^{2}_{2}, \quad \eta^{12}_{12}, \quad \eta^{3}_{3} \quad \cdots \quad \eta^{234}_{234}, \quad \eta^{1234}_{1234},$$

any bidirected graph defines a sub-model by setting parameters to zero (Lupparelli et al, 2009).

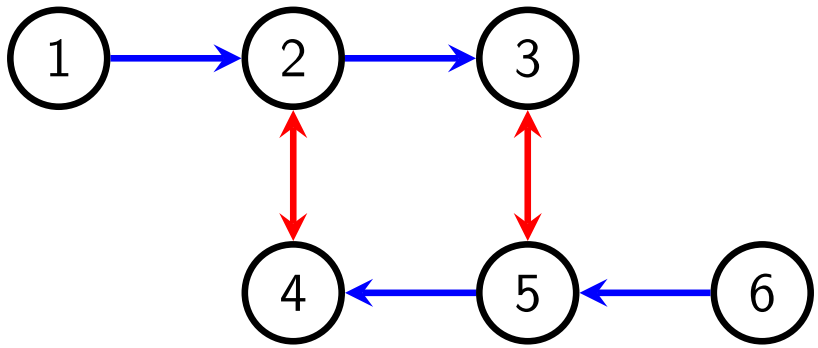


Other Graphical Models

Type IV Chain Graphs. (See Marchetti and Lupparelli, 2011)



Acyclic Directed Mixed Graphs.



Generalise DAGs, bidirected graphs and type IV chain graphs. See Evans and Richardson (2011).

Evaluating MLLPs

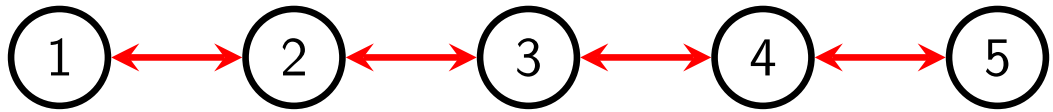
There is no closed-form map from $\boldsymbol{\eta}$ to probabilities $\mathbf{p}$.

Evans and Forcina (2011) provide algorithms to

- evaluate MLLPs;
- maximise likelihood under linear constraints on MLLPs;
- maximise L_1-penalised likelihood with respect to MLLPs.

Simulation

Take $N = 1000$ binary distributions obeying global Markov property with respect to bidirected 5-chain:



Try to recover model from n observations for $n = 10^3$ up to 10^6.

Parametrise saturated model using all margins:

$$\eta_1^1, \quad \eta_2^2, \quad \eta_{12}^{12}, \quad \eta_3^3, \quad \eta_{13}^{13}, \quad \dots \quad \eta_{2345}^{2345}, \quad \eta_{12345}^{12345}.$$

Model corresponds to

$$\eta_{13}^{13} = \eta_{14}^{14} = \eta_{24}^{24} = \eta_{124}^{124} = \eta_{134}^{134} = \dots = \eta_{245}^{245} = \eta_{1245}^{1245} = \eta_{1345}^{1345} = 0$$

(16 constraints)

Simulation Detail

Use $\gamma = 1$ with 10-fold cross-validation to estimate penalty.

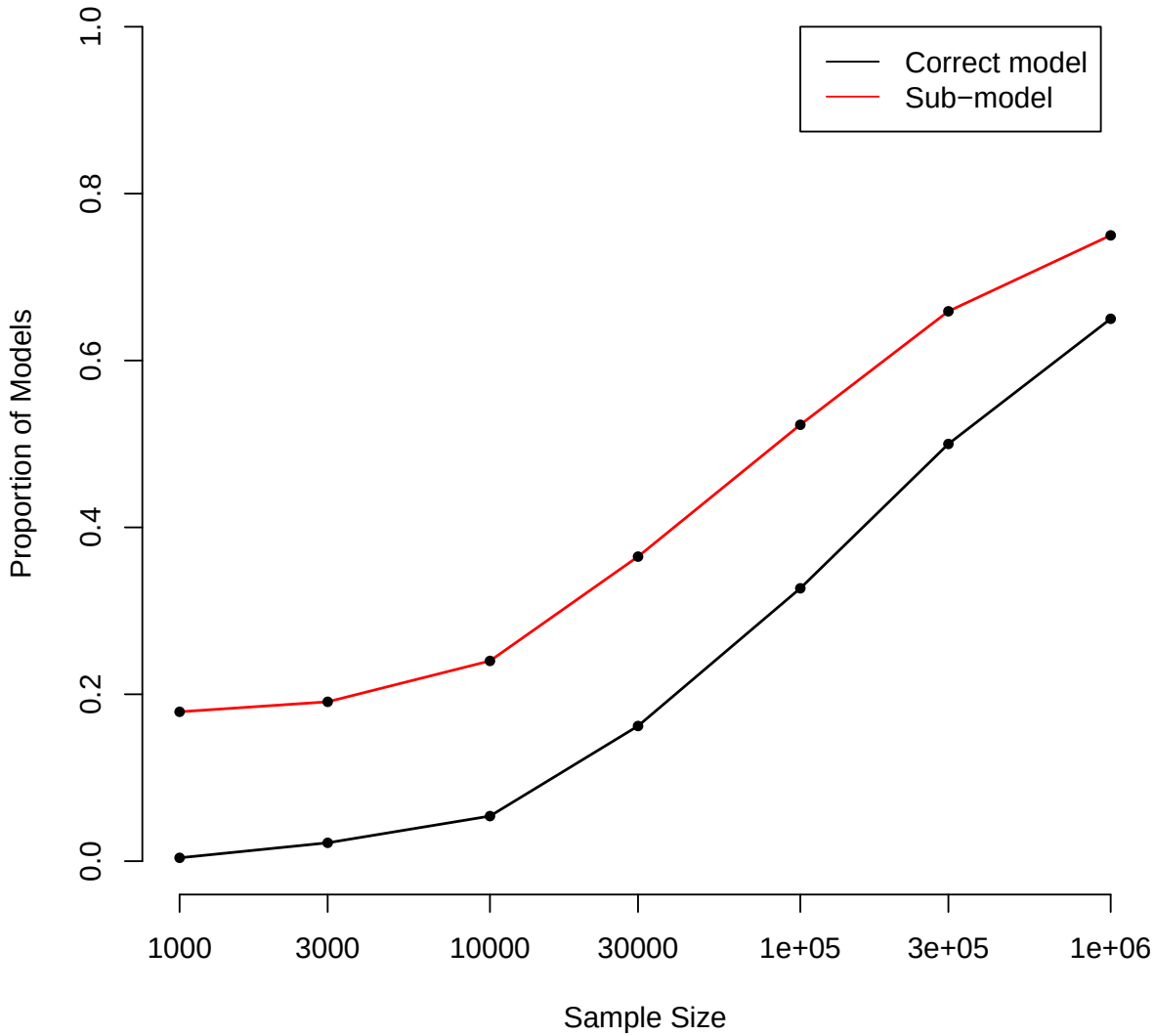
One-way margins are left unconstrained.

Note: correct model requires finding that $\eta_{12345}^{12345} \neq 0$; not easy!

$$\eta_{12345}^{12345} = \frac{1}{32} \log \frac{p_{00000}\, p_{11000}\, p_{10100} \cdots p_{01111}}{p_{10000}\, p_{01000}\, p_{00100} \cdots p_{11111}}.$$

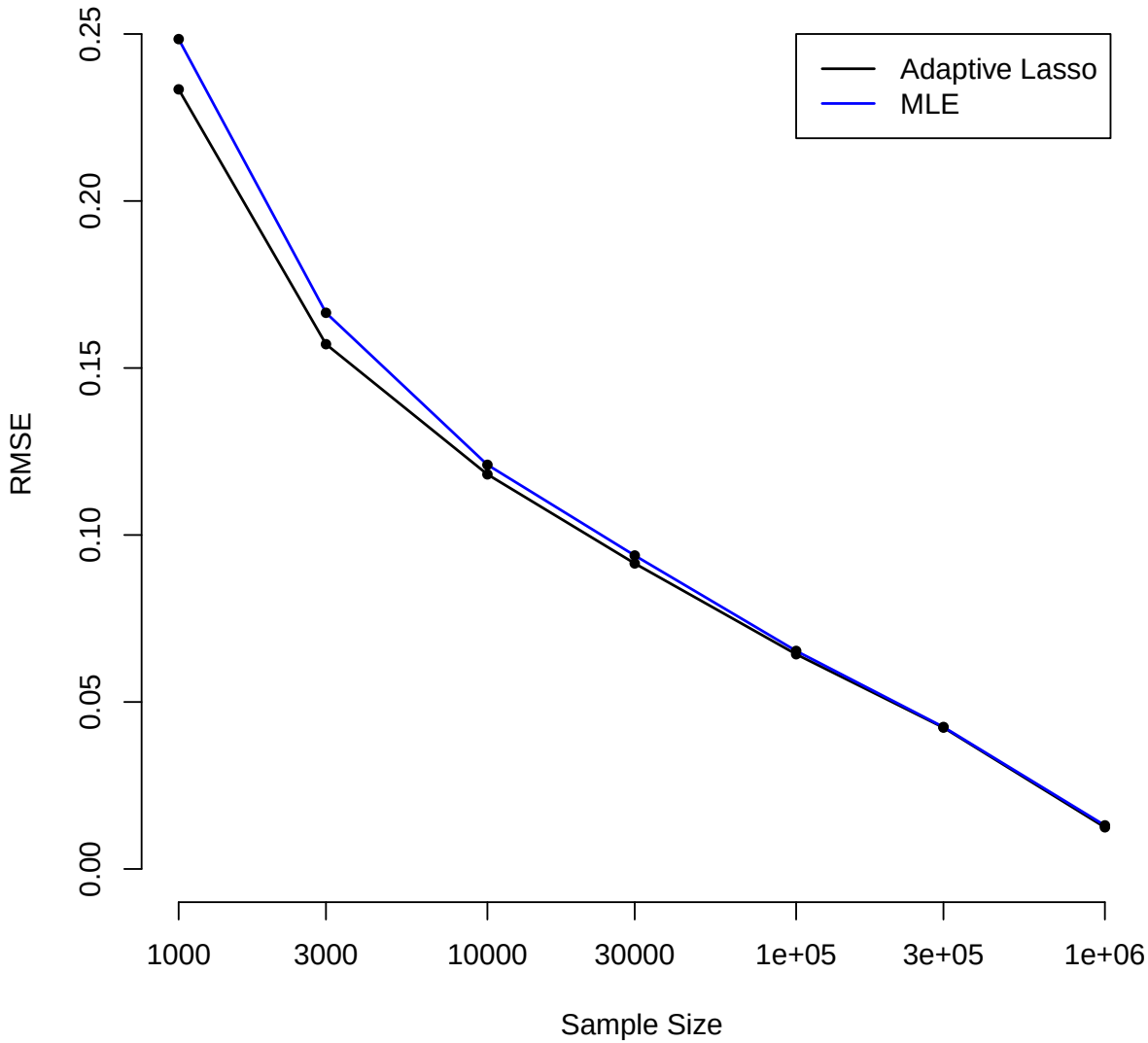
Model Consistency

Proportion of times correct model (or sub-model of correct model) recovered.



Estimation Error

Root mean squared error for estimation of η.



Pros and Cons

This approach doesn't require lots of sub-models to be fitted separately (as e.g. BIC).

Tends to get rid of higher-order interaction terms which may not be useful.

'Oracle' property is misleading — convergence is not uniform, standard errors may be too small.

ν_n must be chosen e.g. by cross-validation in practice.

Problems arise with zero observed counts. One solution: set higher-order params to zero until an MLE exists.

Model returned is not necessarily graphical.

Summary

We have

- seen that the adaptive lasso is model-consistent (and oracle) for regular parametric models;
- given a brief overview of marginal log-linear parameters and their application to discrete graphical models;
- shown how to use the adaptive lasso with these parametrisations to perform model selection within graphical model classes.

Thank you!

References

Bergsma and Rudas (2002), Marginal models for categorical data, *Ann. Stat.*

Colombi and Forcina (2001), Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*.

Evans (2011), Discrete model selection with the adaptive lasso, *draft*.

Evans and Forcina (2011), Two algorithms for fitting constrained marginal models, *submitted*.

Evans and Richardson (2011), Marginal log-linear parameters for graphical Markov models, *submitted*.

Lupparelli, Marchetti, Bergsma (2009), Parameterizations and Fitting of Bi-directed Graph Models to Categorical Data, *Scan. J. Stat*.

Marchetti and Lupparelli (2011), Chain graph models of multivariate regression type for categorical data, *Bernoulli*.

Tibshirani (1996), Regression shrinkage and selection via the lasso, *JRSSB*.

Zou (2006), The adaptive lasso and its oracle properties, *JASA*.

Regularity Conditions

1. For every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}'} \implies \boldsymbol{\theta} = \boldsymbol{\theta}'$.
2. For some σ-finite measure μ, each $P_{\boldsymbol{\theta}}$ has density $p_{\boldsymbol{\theta}}$ with respect to μ.
3. The support $\{x \,|\, p_{\boldsymbol{\theta}}(x) > 0\}$ is independent of $\boldsymbol{\theta}$.
4. For every $\boldsymbol{\theta} \in \Theta$ the log-likelihood $l_n(\boldsymbol{\theta})$ is twice differentiable, its first derivative $\dot{\mathbf{l}}_n(\boldsymbol{\theta})$ (the score) satisfies

   $$\mathbb{E}_{\boldsymbol{\theta}} \dot{\mathbf{l}}_1(\boldsymbol{\theta}) = 0, \qquad \mathbb{E}_{\boldsymbol{\theta}} \dot{\mathbf{l}}_1(\boldsymbol{\theta})^2 < \infty,$$

   and the Fisher information matrix

   $$I(\boldsymbol{\theta}) \equiv -\mathbb{E}_{\boldsymbol{\theta}} \ddot{\mathbf{l}}_1(\boldsymbol{\theta})$$

   is positive definite.
5. The third partial derivatives of $l_1(\boldsymbol{\theta})$ exist and are bounded in expectation in a neighbourhood of the truth $\boldsymbol{\theta}^*$.