Algorithmic approaches to fitting ERG models

Ruth Hummel, Penn State University
Mark Handcock, University of Washington
David Hunter, Penn State University

Research funded by Office of Naval Research Award No. N00014-08-1-1015

MURI meeting, April 24, 2009

Outline

The class of Exponential-family Random Graph Models (ERGMs):

Definition

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{y})\}}{\kappa(\boldsymbol{\eta})}$$

where

- $\mathbf{Y}$ is a random network written as an adjacency matrix so that Y_{ij} is the indicator of an edge from i to j;
- $\mathbf{g}(\mathbf{y})$ is a vector of the network statistics of interest;
- $\boldsymbol{\eta}$ is a vector of parameters corresponding to the vector $\mathbf{g}(\mathbf{y})$;
- $\kappa(\boldsymbol{\eta})$ is the constant of proportionality which makes the probabilities sum to one; intractable

Loglikelihood

The loglikelihood for this class of models is

$$l(\eta) = \eta^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log \sum_{z \in \mathcal{Y}} \exp(\eta^t \mathbf{g}(\mathbf{z})). \quad (1)$$

which can be written:

$$l(\eta) - l(\eta_0) = (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log E_{\eta_0} \left[\exp \left\{ (\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}) \right\} \right], \quad (2)$$

Maximum Pseudolikelihood Estimation

Notation: For a network y and a pair (i,j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i,j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

- Assume no dependence among the Y_{ij}.
- In other words, assume $P(Y_{ij} = 1) = P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$.
- Then some algebra gives

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \left[g(y_{ij}^+) - g(y_{ij}^-) \right],$$

so θ is estimated by straightforward logistic regression.

- Result: The maximum pseudolikelihood estimate.

Maximum Pseudolikelihood Estimation

Notation: For a network y and a pair (i,j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i,j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

- Assume no dependence among the Y_{ij}.
- In other words, assume $P(Y_{ij} = 1) = P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$.
- Then some algebra gives

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \left[g(y_{ij}^+) - g(y_{ij}^-) \right],$$

  so θ is estimated by straightforward logistic regression.
- Result: The maximum pseudolikelihood estimate.

MPLE's behavior

- MLE (maximum likelihood estimation): Well-established method but very hard because the normalizing constant $\kappa(\alpha)$ is difficult (usually impossible) to evaluate, so we approximate it instead.
- MPLE (maximum pseudo-likelihood estimation): Easy to do using logistic regression, but based on an independence assumption that is often not justified.

Several authors, notably van Duijn et al. (2009), argue forcefully against the use of MPLE.

Back to the loglikelihood

- Remember that the loglikelihood

$$l(\eta) = \eta^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log \sum_{z \in \mathcal{Y}} \exp(\eta^t \mathbf{g}(\mathbf{z})). \tag{3}$$

can be written:

$$l(\eta) - l(\eta_0) = (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log E_{\eta_0} \left[\exp \left\{ (\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}) \right\} \right],$$

This leads to our first approximation for the MLE:

- MCMC MLE idea: Pick θ_0, draw $Y_1, \ldots, Y_m$ from this model using MCMC, then approximate the population mean above by a sample mean.
- We can take θ_0 to be, for example, the MPLE.

Back to the loglikelihood

- Remember that the loglikelihood

$$l(\eta) = \eta^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log \sum_{z \in \mathcal{Y}} \exp(\eta^t \mathbf{g}(\mathbf{z})). \tag{3}$$

  can be written:

$$l(\eta) - l(\eta_0) = (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log E_{\eta_0} \left[\exp \left\{ (\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}) \right\} \right],$$

  This leads to our first approximation for the MLE:

- MCMC MLE idea: Pick θ_0, draw $Y_1, \ldots, Y_m$ from this model using MCMC, then approximate the population mean above by a sample mean.
- We can take θ_0 to be, for example, the MPLE.

Back to the loglikelihood

- Remember that the loglikelihood

$$l(\eta) = \eta^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log \sum_{z \in \mathcal{Y}} \exp(\eta^t \mathbf{g}(\mathbf{z})). \tag{3}$$

  can be written:

$$l(\eta) - l(\eta_0) = (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log E_{\eta_0} \left[\exp \left\{ (\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}) \right\} \right],$$

  This leads to our first approximation for the MLE:
- MCMC MLE idea: Pick θ_0, draw $Y_1, \ldots, Y_m$ from this model using MCMC, then approximate the population mean above by a sample mean.
- We can take θ_0 to be, for example, the MPLE.

Existence of MLE?

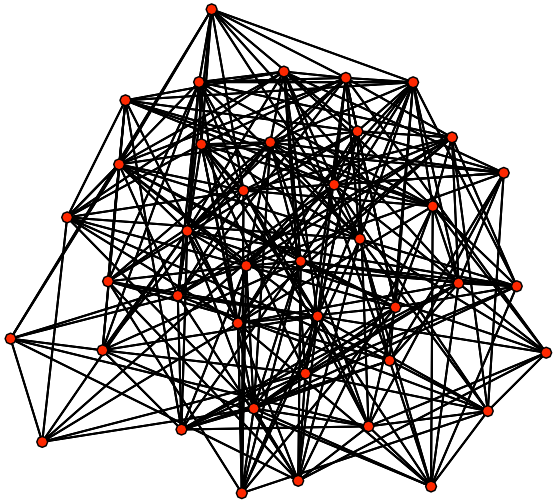
There are two issues we need to be concerned with when using MCMC MLE:

- First, if the observed $\mathbf{g}(\mathbf{y}_{\text{obs}})$ is not in the interior of the convex hull of the sampled (MCMC generated) $\mathbf{g}(\mathbf{y}_i)$, then no maximizer of the approximate loglikelihood ratio exists.
- Also, the approximation of the likelihood surface,

$$l(\eta) - l(\eta_0) \approx (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{obs}) - \log \frac{1}{m} \sum_{i=1}^{m} \exp((\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}_i)),$$

  is not very good when we get far from η_0.

Quick Erdős-Rényi illustration

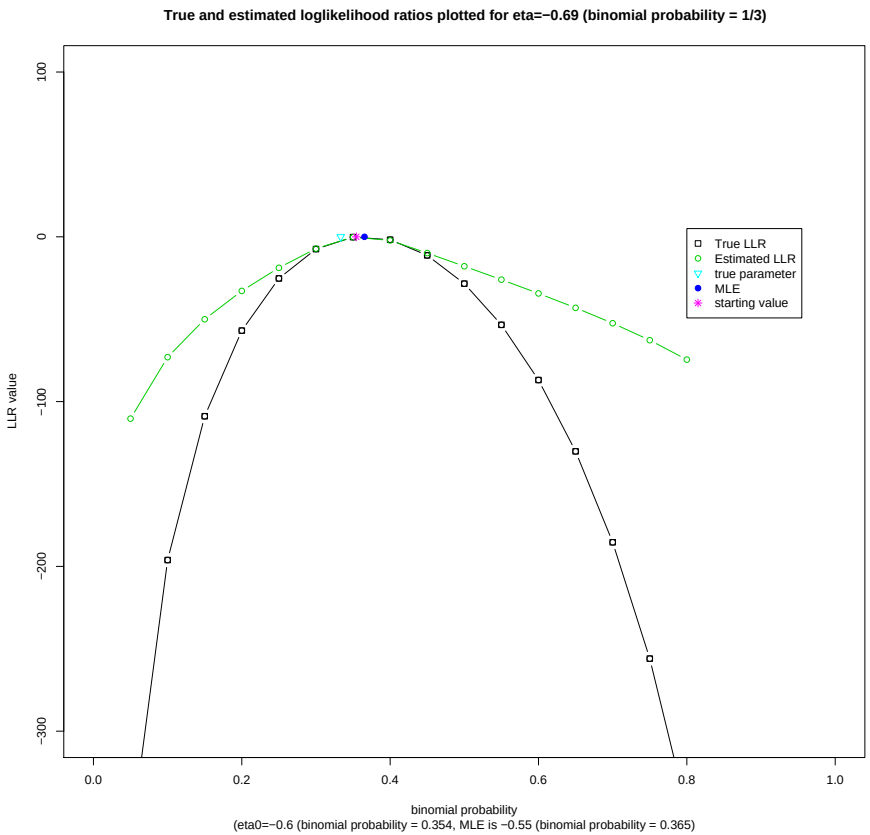


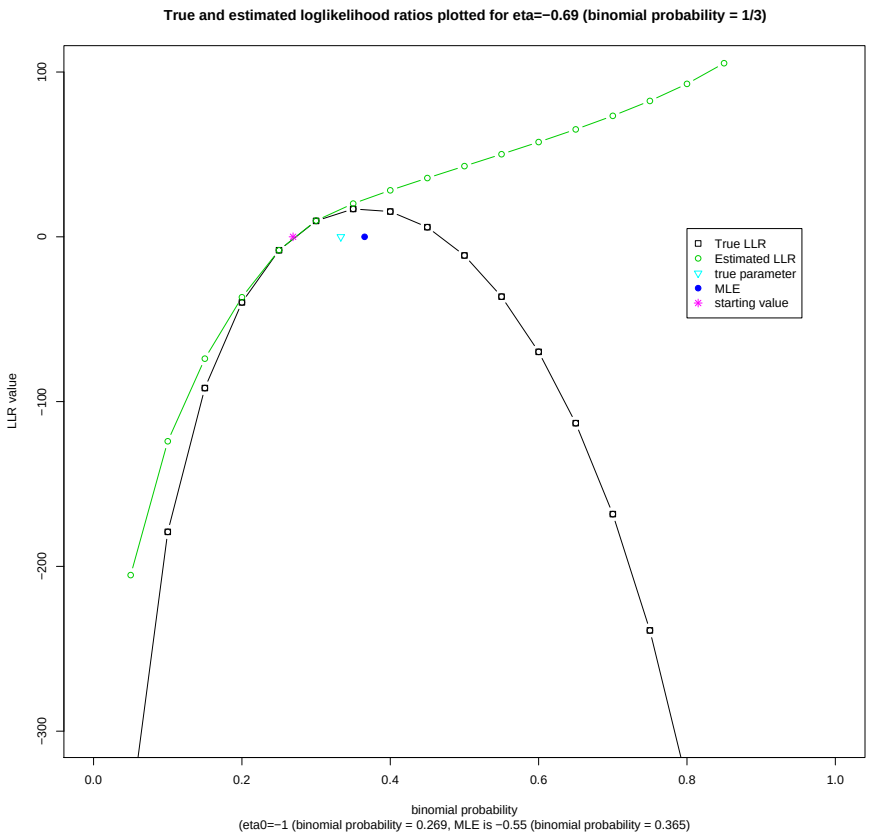
The Erdős-Rényi model can be written as an ERGM, if $\mathbf{g}(\mathbf{y})$ is the number of edges of $\mathbf{y}$.

Since each edge exists independently with probability b,

$$P(\mathbf{Y} = \mathbf{y}) = (b)^{g(\mathbf{y})} (1-b)^{n^* - g(\mathbf{y})} = \left(\frac{b}{1-b}\right)^{g(\mathbf{y})} (1-b)^{n^*}.$$

Comparison of the true loglikelihood ratio and its estimation for an Erdos-Renyi graph





Outline

Mean value parametrization

1. We would like to move our arbitrary initial value η_o close enough to the MLE that samples (of the statistics) generated from η_0 cover the observed statistics, $\mathbf{g}(\mathbf{y}_{obs})$.
2. New Idea: Intuitively, then, we might try to find a way to make the journey from η_0 to the MLE in pieces.

In order to take steps toward the MLE, we need to have some idea where we are going. We obviously don't know where the MLE is, but we do know that the MLE has the following property:

Definition

The MLE, if it exists, is the unique parameter vector η satisfying $E_\eta \mathbf{g}(\mathbf{Y}) = \mathbf{g}(\mathbf{y}_{\text{obs}}) = \hat{\xi}$.

This is true for any exponential family (Barndorff-Nielson, 1978, and Brown, 1986).

Mean value parametrization

1. We would like to move our arbitrary initial value η_o close enough to the MLE that samples (of the statistics) generated from η_0 cover the observed statistics, $\mathbf{g}(\mathbf{y}_{obs})$.
2. New Idea: Intuitively, then, we might try to find a way to make the journey from $\boldsymbol{\eta}_0$ to the MLE in pieces.

In order to take steps toward the MLE, we need to have some idea where we are going. We obviously don't know where the MLE is, but we do know that the MLE has the following property:

Definition

The MLE, if it exists, is the unique parameter vector $\boldsymbol{\eta}$ satisfying $E_{\boldsymbol{\eta}}\mathbf{g}(\mathbf{Y}) = \mathbf{g}(\mathbf{y}_{\text{obs}}) = \hat{\boldsymbol{\xi}}$.

This is true for any exponential family (Barndorff-Nielson, 1978, and Brown, 1986).

Mean value parametrization

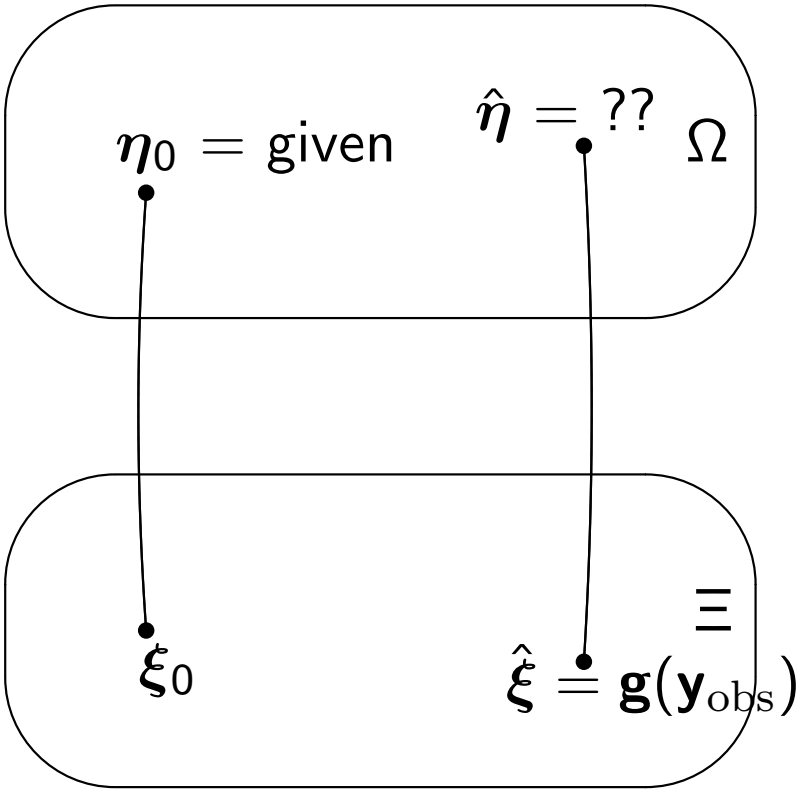
1. We would like to move our arbitrary initial value η_o close enough to the MLE that samples (of the statistics) generated from η_0 cover the observed statistics, $\mathbf{g}(\mathbf{y}_{obs})$.
2. New Idea: Intuitively, then, we might try to find a way to make the journey from $\boldsymbol{\eta}_0$ to the MLE in pieces.

In order to take steps toward the MLE, we need to have some idea where we are going. We obviously don't know where the MLE is, but we do know that the MLE has the following property:

Definition

The MLE, if it exists, is the unique parameter vector $\boldsymbol{\eta}$ satisfying $E_{\boldsymbol{\eta}}\mathbf{g}(\mathbf{Y}) = \mathbf{g}(\mathbf{y}_{\text{obs}}) = \hat{\boldsymbol{\xi}}$.

This is true for any exponential family (Barndorff-Nielson, 1978, and Brown, 1986).



- Ω is the original η parameter space (defined for the change statistics)
- Ξ is the corresponding mean value parameter space
- η_0 is the (given) initial value of η in the Markov Chain
- ξ_0 is the vector of mean statistics corresponding to η_0
- $\hat{\eta}$ is the (unknown) MLE
- $\hat{\xi}$ is the observed statistics, $\mathbf{g}(\mathbf{y}_{\text{obs}})$, which is the corresponding mean statistics vector for $\hat{\eta}$.

Taking steps

- We now want to move toward $\hat{\boldsymbol{\xi}} = \mathbf{g}(\mathbf{y}_{\text{obs}})$ in mean value parameter space. Here we specify a step length, $0 \leq \gamma \leq 1$, as a fraction of the distance toward $\hat{\boldsymbol{\xi}}$ that we want to traverse. We move the fraction γ toward $\hat{\boldsymbol{\xi}}$ and call this point $\boldsymbol{\xi}_1 = \gamma\hat{\boldsymbol{\xi}} + (1-\gamma)\hat{\boldsymbol{\xi}}_0$.
- At each step, we choose this fraction to be the biggest move toward $\mathbf{g}(\mathbf{y}_{\text{obs}})$ that does not leave the convex hull of the current sample.

Taking steps

- We now want to move toward $\hat{\boldsymbol{\xi}} = \mathbf{g}(\mathbf{y}_{\text{obs}})$ in mean value parameter space. Here we specify a step length, $0 \leq \gamma \leq 1$, as a fraction of the distance toward $\hat{\boldsymbol{\xi}}$ that we want to traverse. We move the fraction γ toward $\hat{\boldsymbol{\xi}}$ and call this point $\boldsymbol{\xi}_1 = \gamma\hat{\boldsymbol{\xi}} + (1-\gamma)\hat{\boldsymbol{\xi}}_0$.
- At each step, we choose this fraction to be the biggest move toward $\mathbf{g}(\mathbf{y}_{\text{obs}})$ that does not leave the convex hull of the current sample.

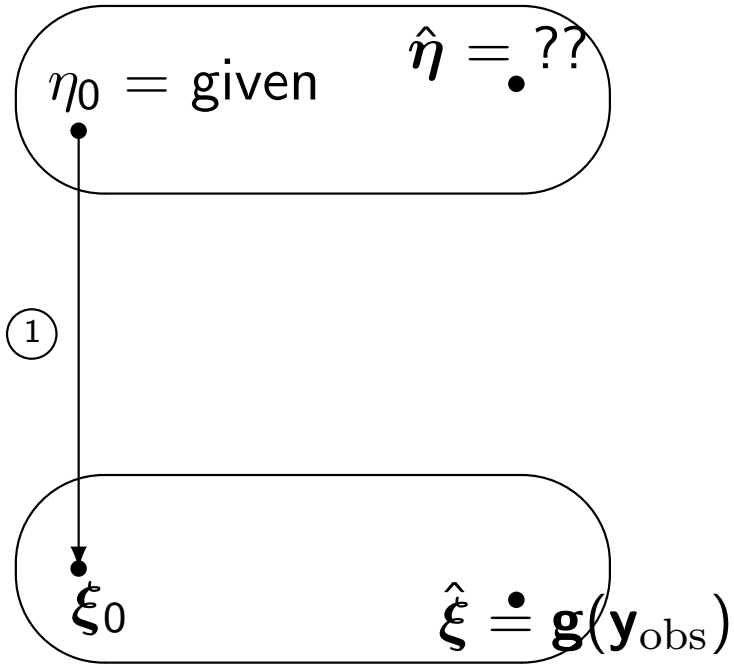
Finding the MCMC MLE

Next, using the sample from the model defined by η_0, we maximize the approximate loglikelihood

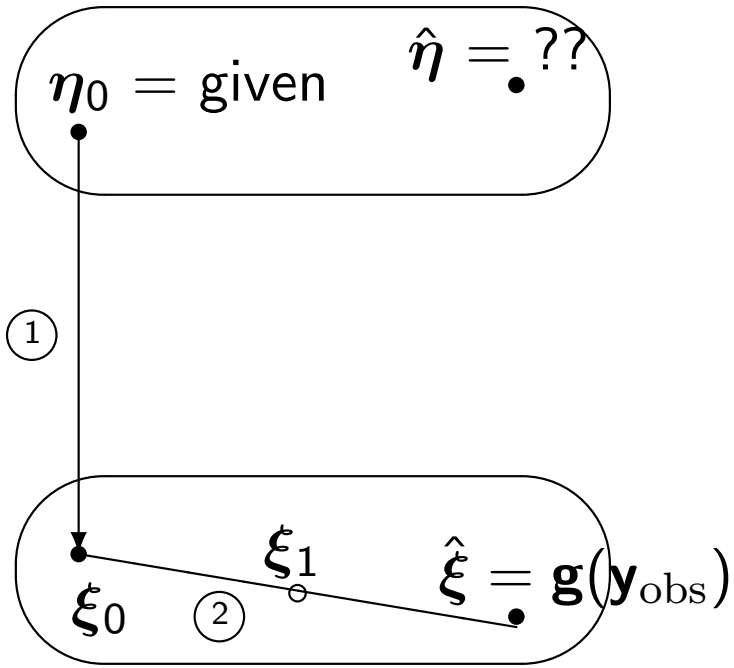
$$l(\eta) - l(\eta_0) \approx (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_{obs}) - \log \frac{1}{m} \sum_{i=1}^{m} \exp((\eta - \eta_0)^t \mathbf{g}(\mathbf{Y}_i)), \quad (6)$$

but with ξ_1 *substituted in place of* $\mathbf{g}(\mathbf{y}_{\mathrm{obs}})$. The resulting maximizer will be called η_1. The process then repeats, with η_1 taking the place of η_0.

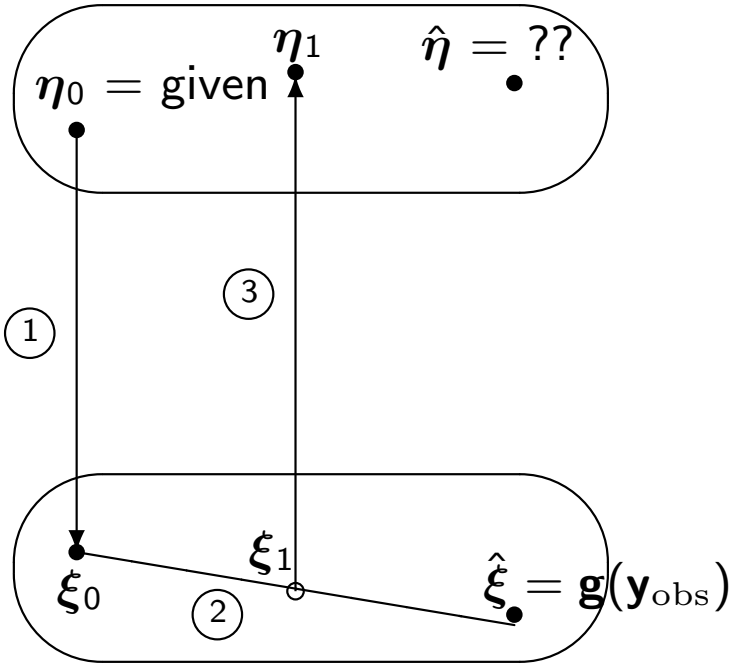
Partial stepping in Mean Value Parameter space



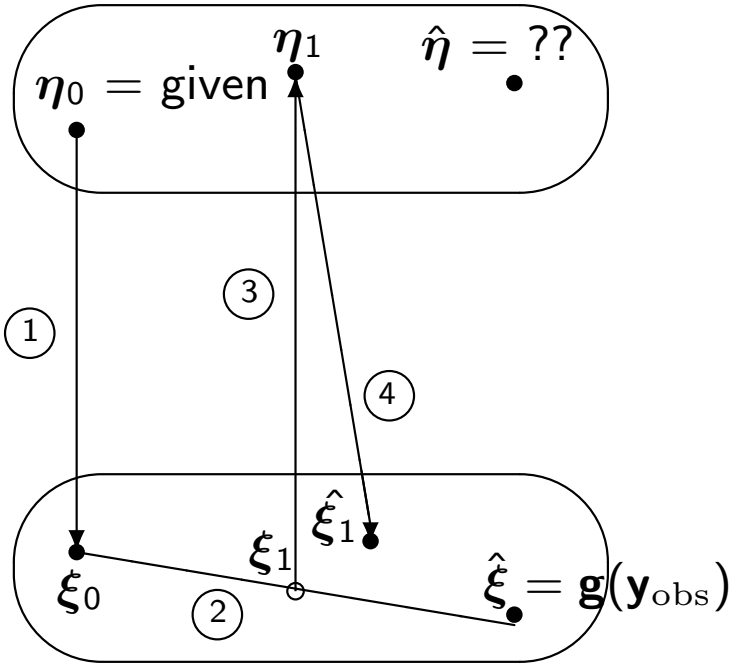
0. Set η_0.
1. Take an MCMC sample from the model defined by $\eta = \eta_0$ to get ξ_0.



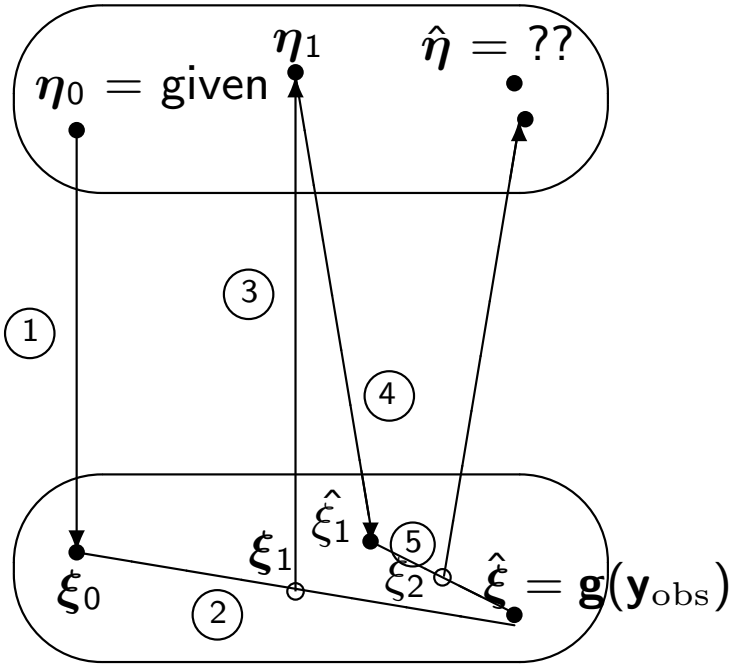
2. Go $\gamma\%$ toward $\hat{\xi} = \mathbf{g}(\mathbf{y}_{\text{obs}})$ in mean value parameter space. Call this ξ_1.



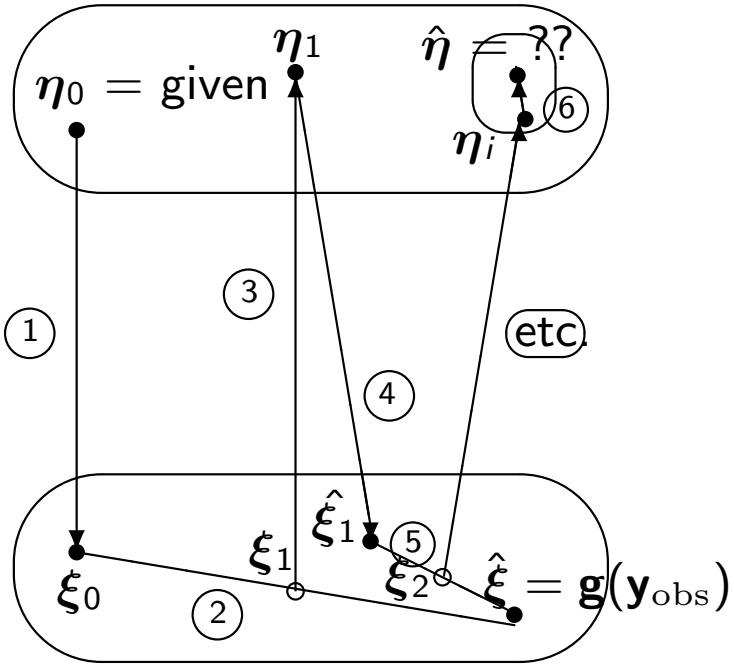
3. Use η_0 in MCMC MLE to find the η that corresponds to ξ_1. Call this η_1.



4. Re-estimate ξ_1 from an MCMC sample from the model defined by $\eta = \eta_1$. Call this $\hat{\xi}_1$.



5. Repeat step (2) by going $\gamma\%$ toward $\hat{\boldsymbol{\xi}} = \mathbf{g}(\mathbf{y}_{\text{obs}})$ from $\hat{\boldsymbol{\xi}}_1$. Call this $\boldsymbol{\xi}_2$. Also repeat steps (3) and (4) to obtain $\hat{\boldsymbol{\xi}}_2$. Keep going until $\mathbf{g}(\mathbf{y}_{obs})$ is in the convex hull of the new sample.



6. Use MCMC MLE with η_i as the initial value to find $\hat{\eta}$. If we have made it into the appropriate neighborhood around $\hat{\eta}$, this will now be possible.

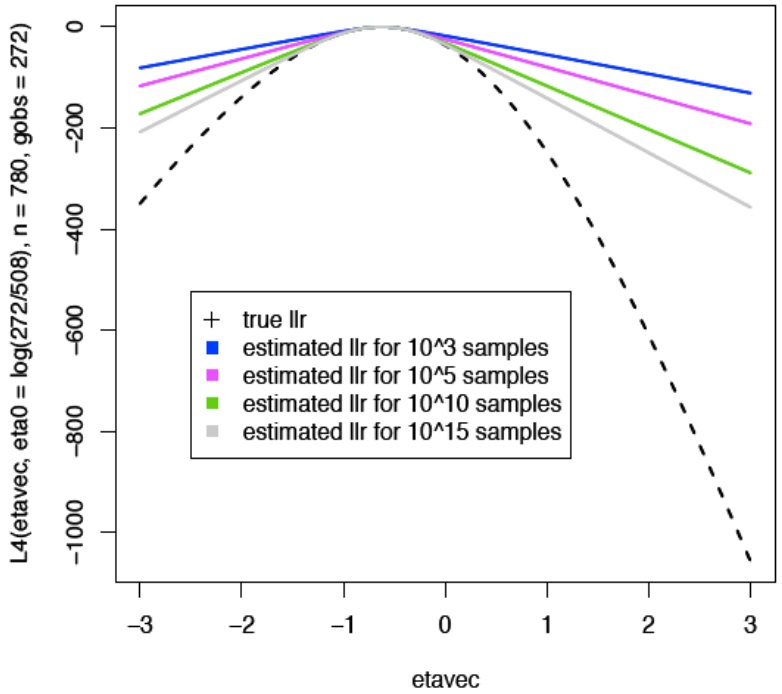
In review:

In other words, for each $t \geq 0$, we first use MCMC to draw a random sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ from the model determined by η_t, then we set

$$\begin{aligned}
\hat{\xi}_t &= \frac{1}{m} \sum_{i=1}^{m} \mathbf{g}(\mathbf{Y}_i); \\
\xi_{t+1} &= \gamma_t \hat{\xi} + (1 - \gamma_t) \hat{\xi}_t; \\
\eta_{t+1} &= \arg\max_{\eta} \left\{ (\eta - \eta_0)^t \xi_{t+1} - \log \left[\frac{1}{m} \sum_{i=1}^{m} \exp\left\{ (\eta - \eta_0)^t \mathbf{g}(\mathbf{y}_i) \right\} \right] \right.
\end{aligned}$$

We iterate until $\hat{\xi} = \mathbf{g}(\mathbf{y}_{\text{obs}})$ is in the convex hull of the statistics generated from $\hat{\xi}_t$.

A second approximation: the lognormal



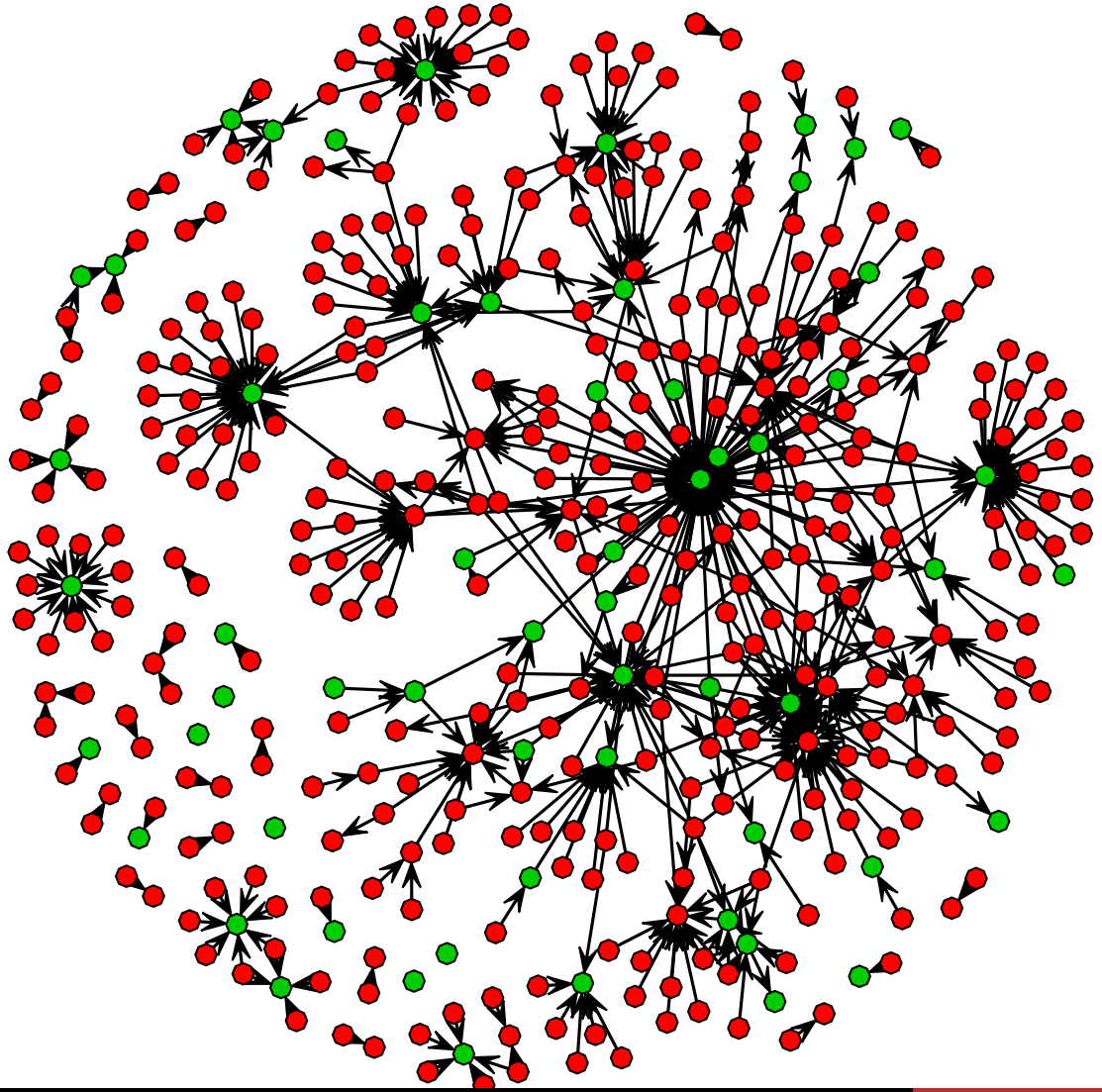
- Here, the ERGM is very simple, $g(y) =$ edges.
- In other words, the model is binomial.

- Nevertheless, naive MCMC approximation of the log likelihood ratio is not good far from θ_0, even for gigantic samples.
- One possible remedy: Assume $(\theta - \theta_0)^t g(Y)$ is normally distributed in

$$\ell(\theta) - \ell(\theta_0) = (\theta - \theta_0)^t g(y^{\text{obs}}) - \log \mathrm{E}_{\theta_0} \left[\exp \left\{ (\theta - \theta_0)^t g(Y) \right\} \right].$$

Outline

Example Bionet: E. Coli (Salgado et al 2001)



- A node is an operon
- Edge $A \longrightarrow B$ means A encodes a transcription factor that regulates B.
- Green indicates self-regulation

Another ERGM for the E. Coli network

We fit a model similar to that of Saul and Filkov (2007)

Term(s)	**Description**:
Edges	Number of Edges
2-Deg, . . . , 5-Deg	Nodes with degree 2, . . . , 5
GWDeg	Single statistic: Weighted sum of 1-Deg, . . . , $(n-1)$-Deg with weights tending to 1 at a geometric rate

```
model <- ergm(ecoli2 ~ edges + degree(2:5) +
gwdegree(0.25, fixed=TRUE), MPLEonly=TRUE)

MPLE
   edges  degree2  degree3  degree4  degree5 gwdegree
   -5.35    -2.58    -3.06    -2.39    -1.85     8.13
```

Another ERGM for the E. Coli network

We fit a model similar to that of Saul and Filkov (2007)

Term(s)	Description:
Edges	Number of Edges
2-Deg, . . . , 5-Deg	Nodes with degree 2, . . . , 5
GWDeg	Single statistic: Weighted sum of 1-Deg, . . . , $(n-1)$-Deg with weights tending to 1 at a geometric rate

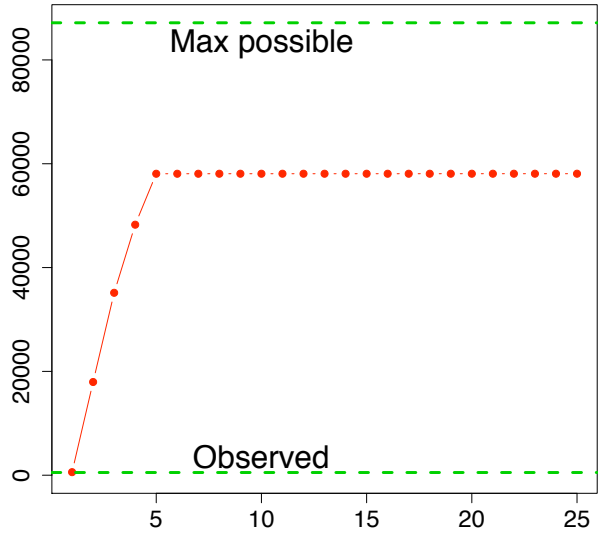
```
model <- ergm(ecoli2 ~ edges + degree(2:5) +
gwdegree(0.25, fixed=TRUE), MPLEonly=TRUE)
```

```
MPLE
   edges  degree2  degree3  degree4  degree5 gwdegree
   -5.35    -2.58    -3.06    -2.39    -1.85     8.13
```

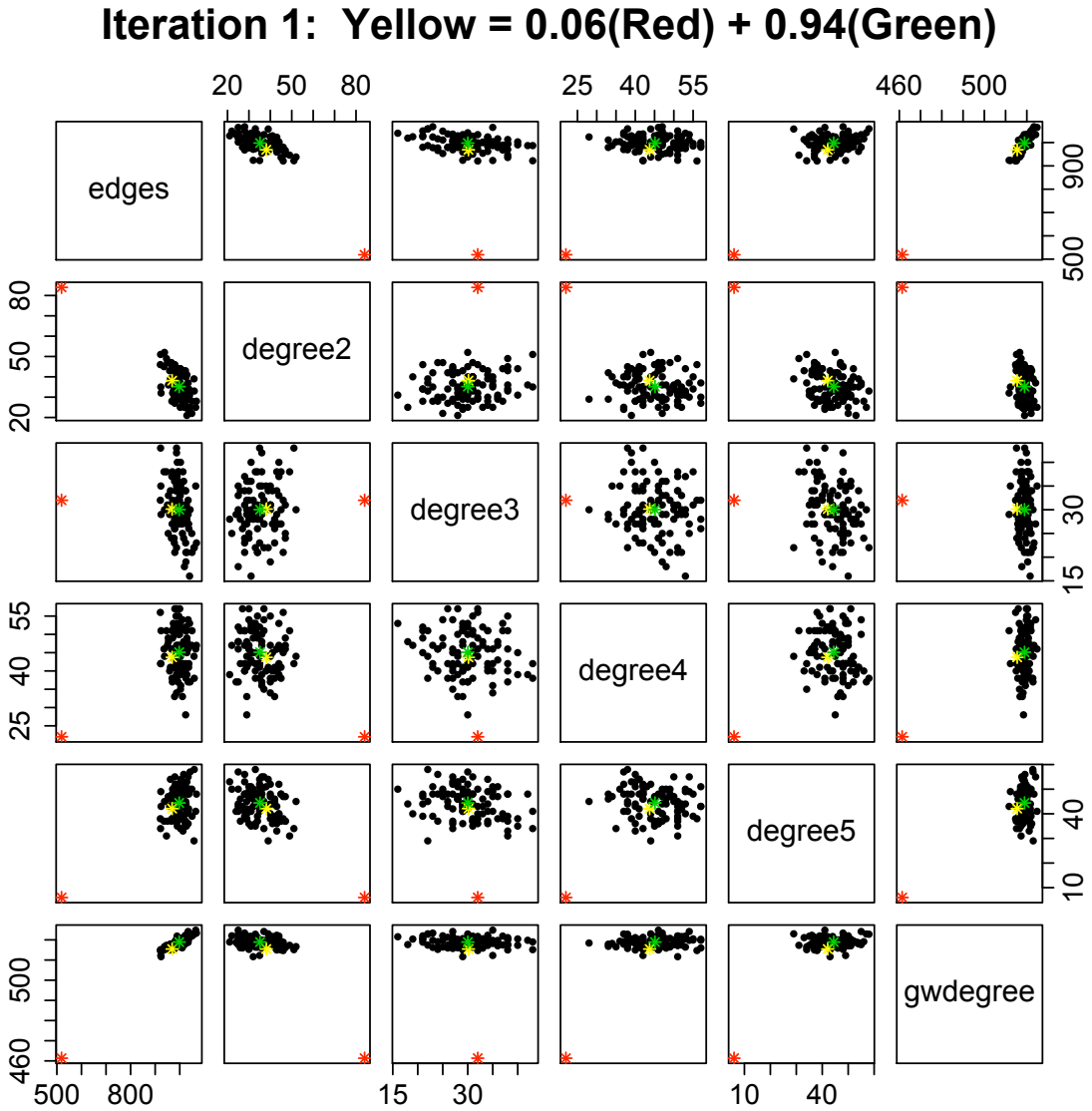
MPLE fit is degenerate

With the MPLE we encounter problems:

Here is a time-series plot of the edge-count of the 25 networks generated from the MPLE:

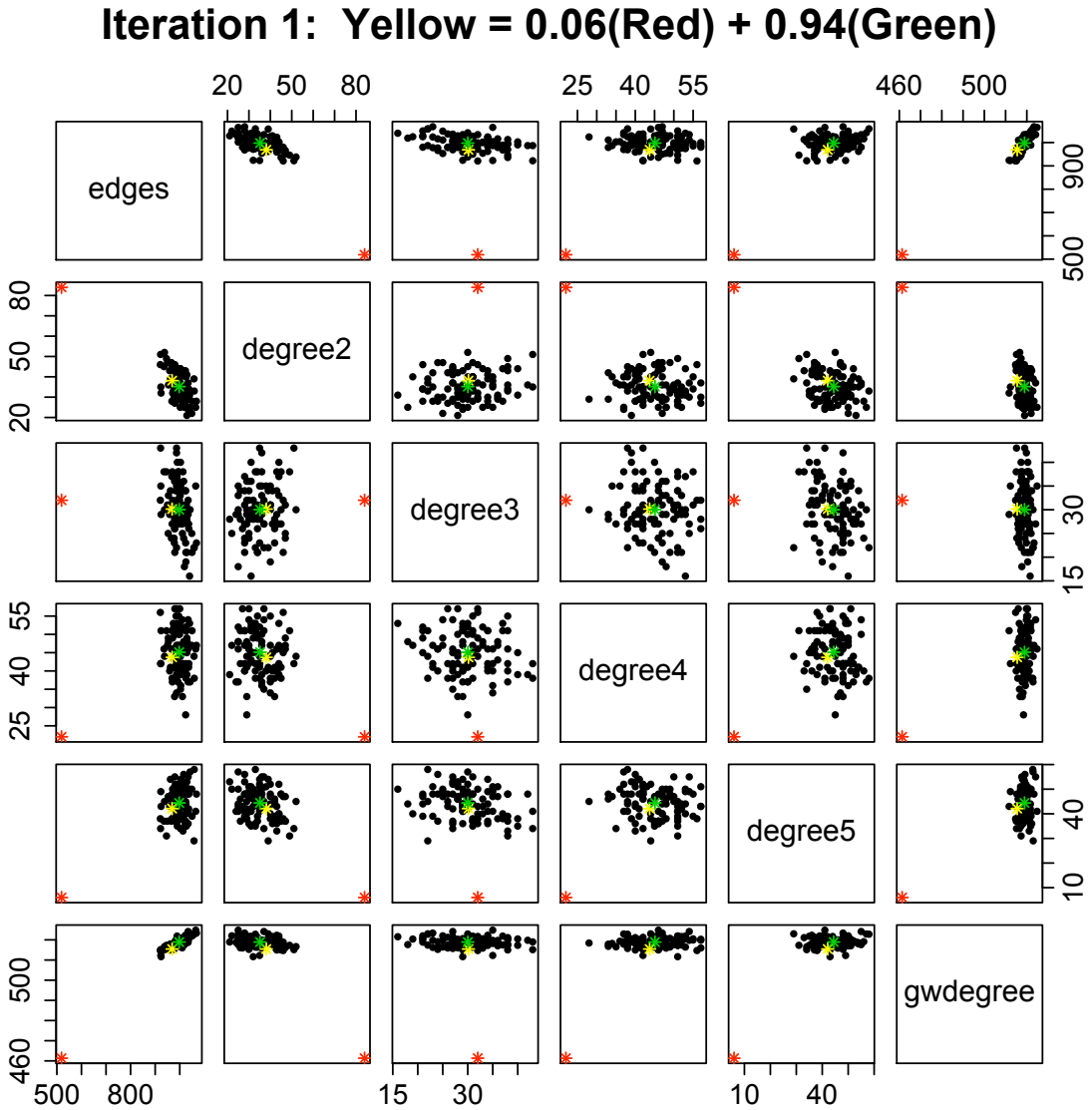


A sample from the MPLE-fitted model



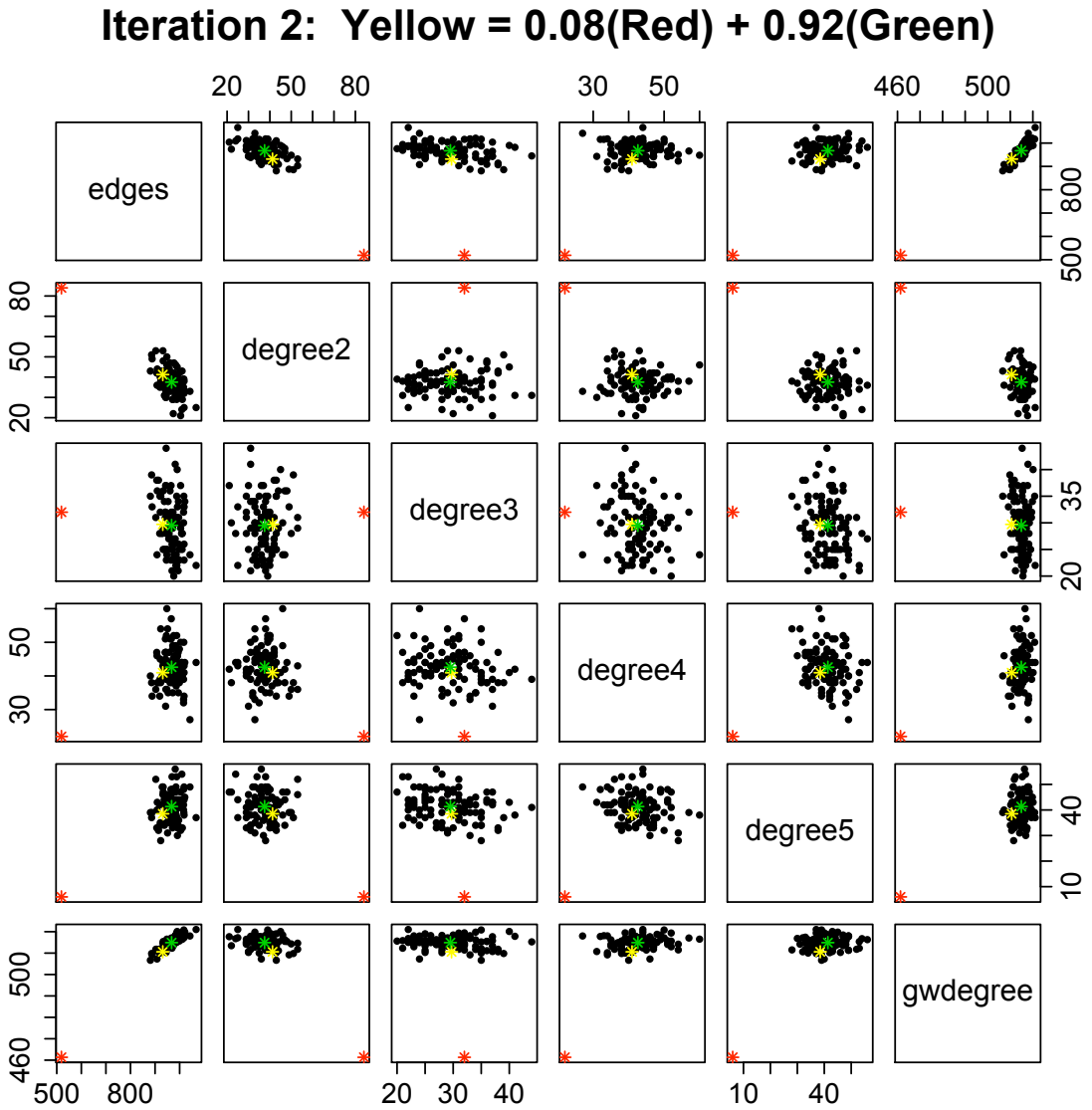
- Red: $g(y^{\text{obs}})$
- Green: Sample mean
- Theory: No maximizer of the approximated likelihood exists because $g(y^{\text{obs}})$ is not in the interior of the convex hull of the sampled points.
- However, the likelihood depends on the data only through $g(y^{\text{obs}})$
- Idea: What if we pretend $g(y^{\text{obs}})$ is the yellow point?

A sample from the MPLE-fitted model



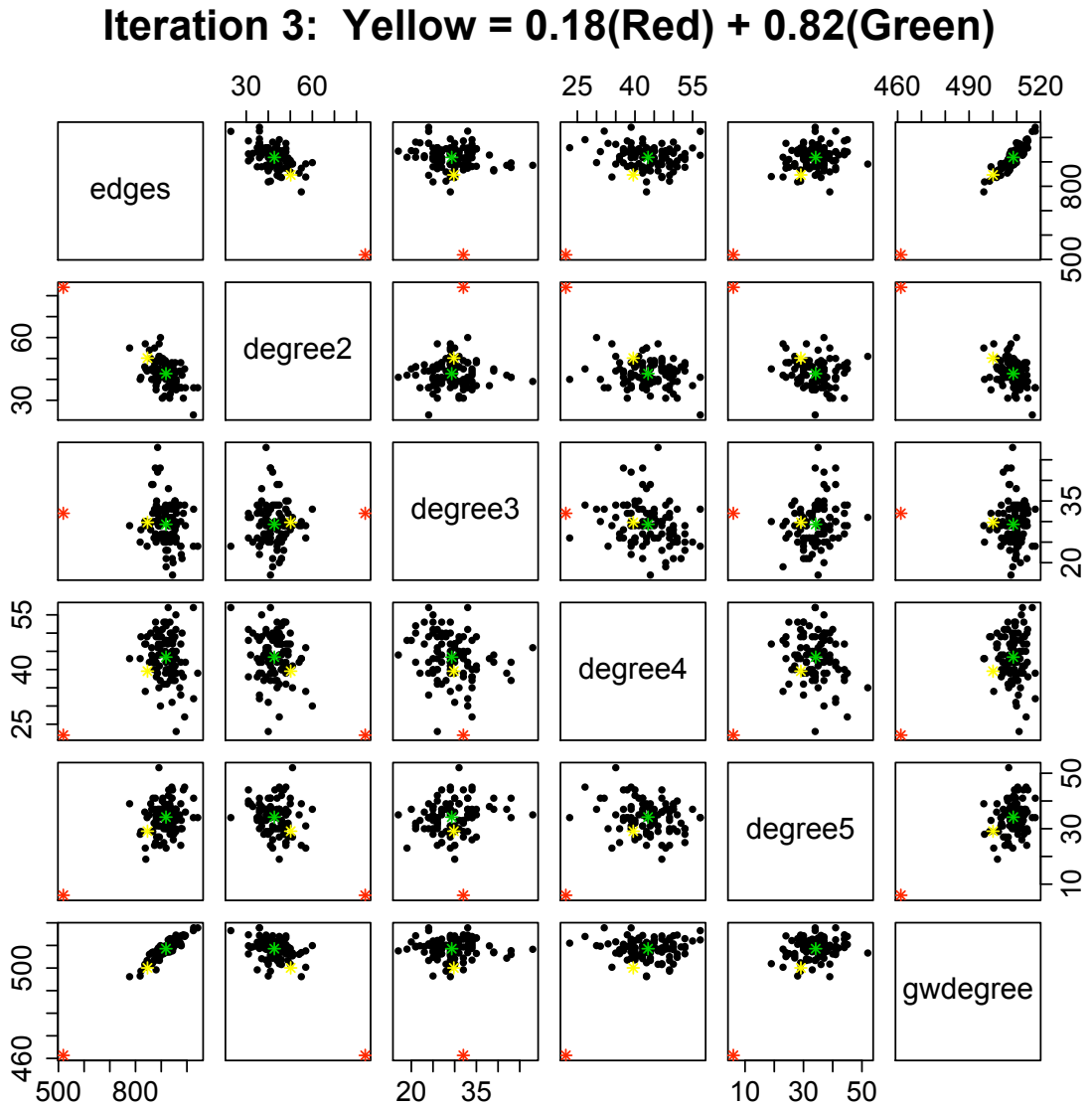
- Red: $g(y^{\text{obs}})$
- Green: Sample mean
- Theory: No maximizer of the approximated likelihood exists because $g(y^{\text{obs}})$ is not in the interior of the convex hull of the sampled points.
- However, the likelihood depends on the data only through $g(y^{\text{obs}})$
- Idea: What if we pretend $g(y^{\text{obs}})$ is the yellow point?

A sample from a model with a better θ_0



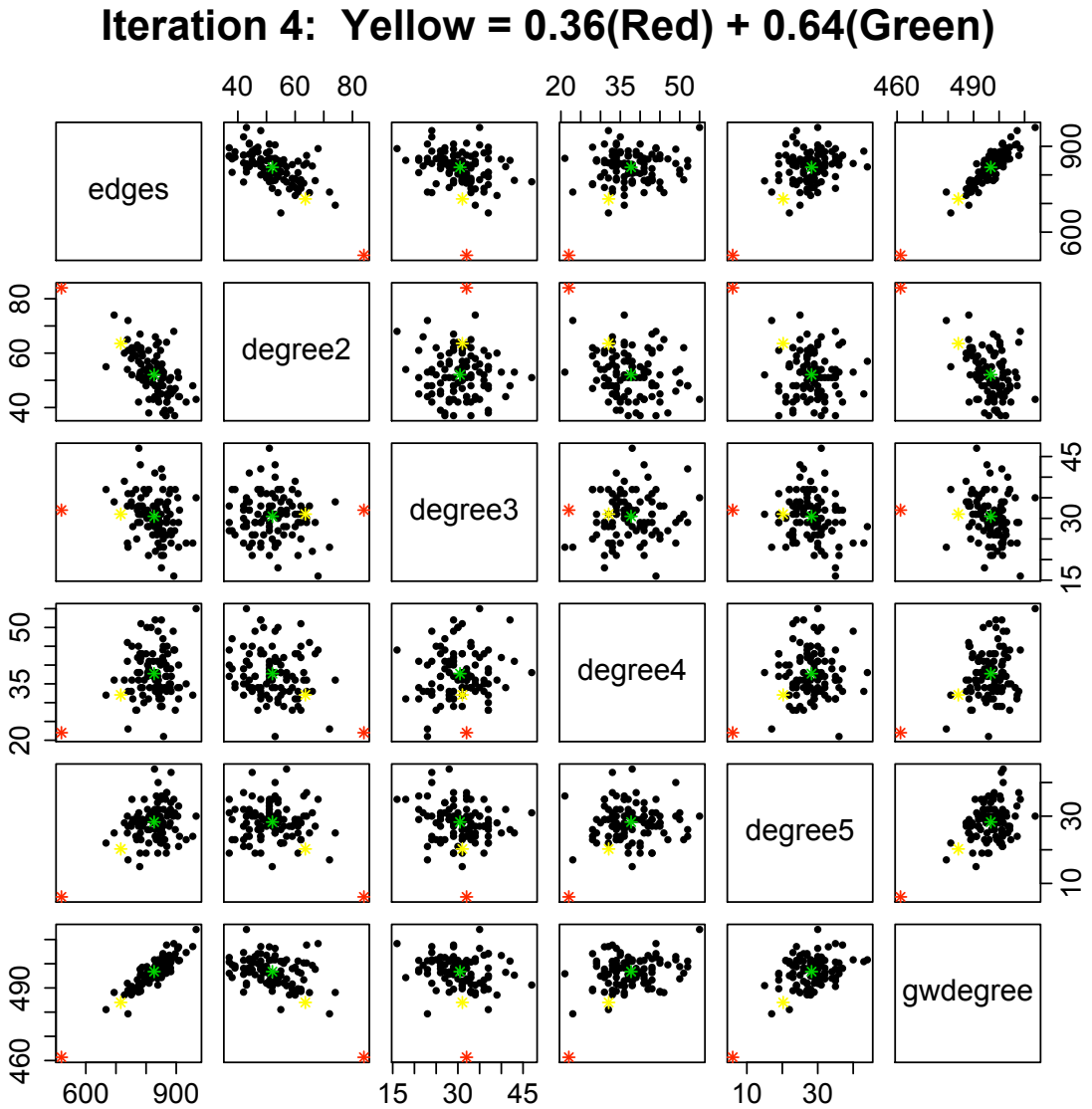
- For θ_0 we have replaced the MPLE by the MCMC "MLE" obtained by pretending that $g(y^{\text{obs}})$ was the yellow point on the previous slide.

A sample from a model with a better θ_0

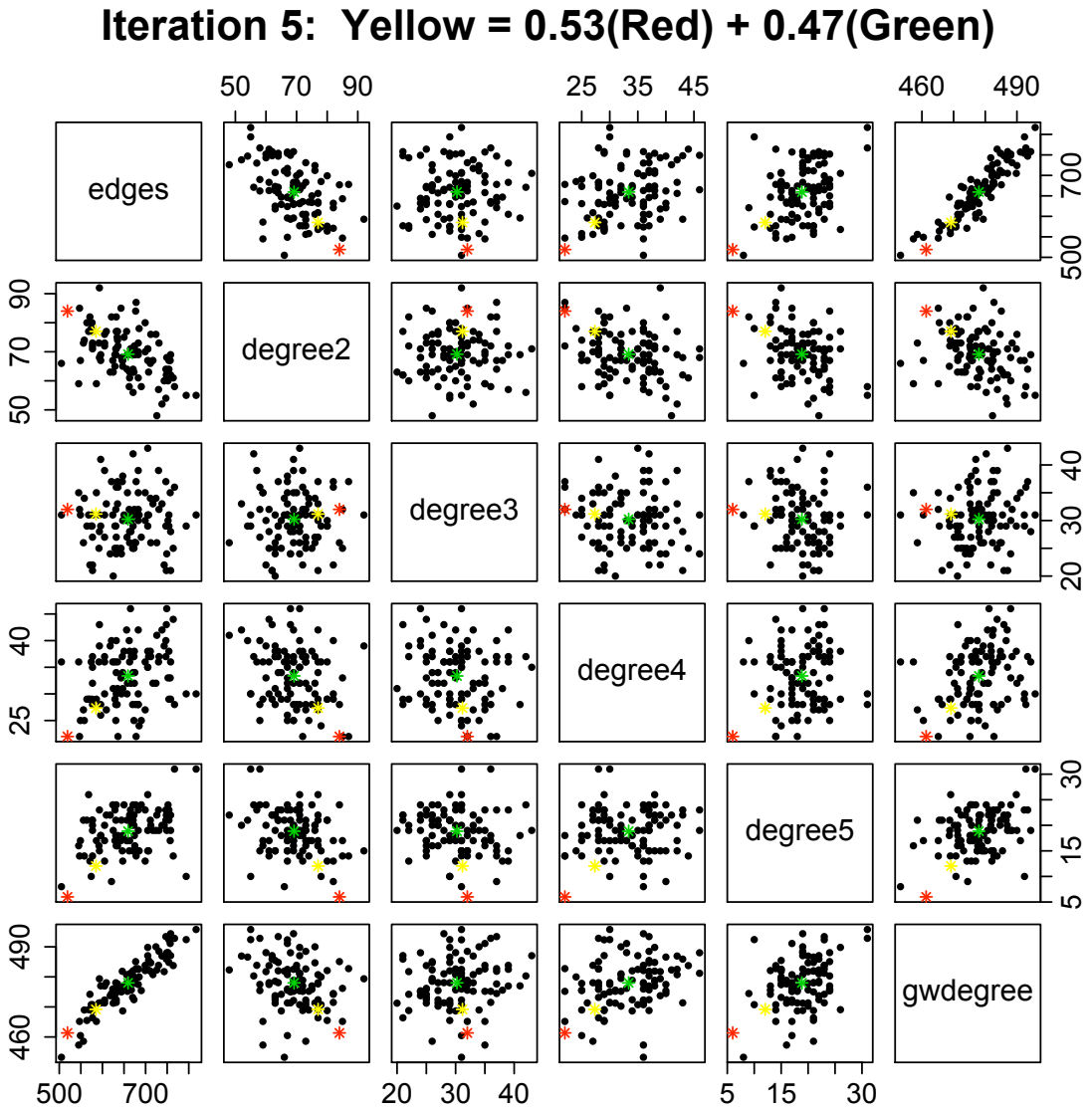


- Continue to iterate...

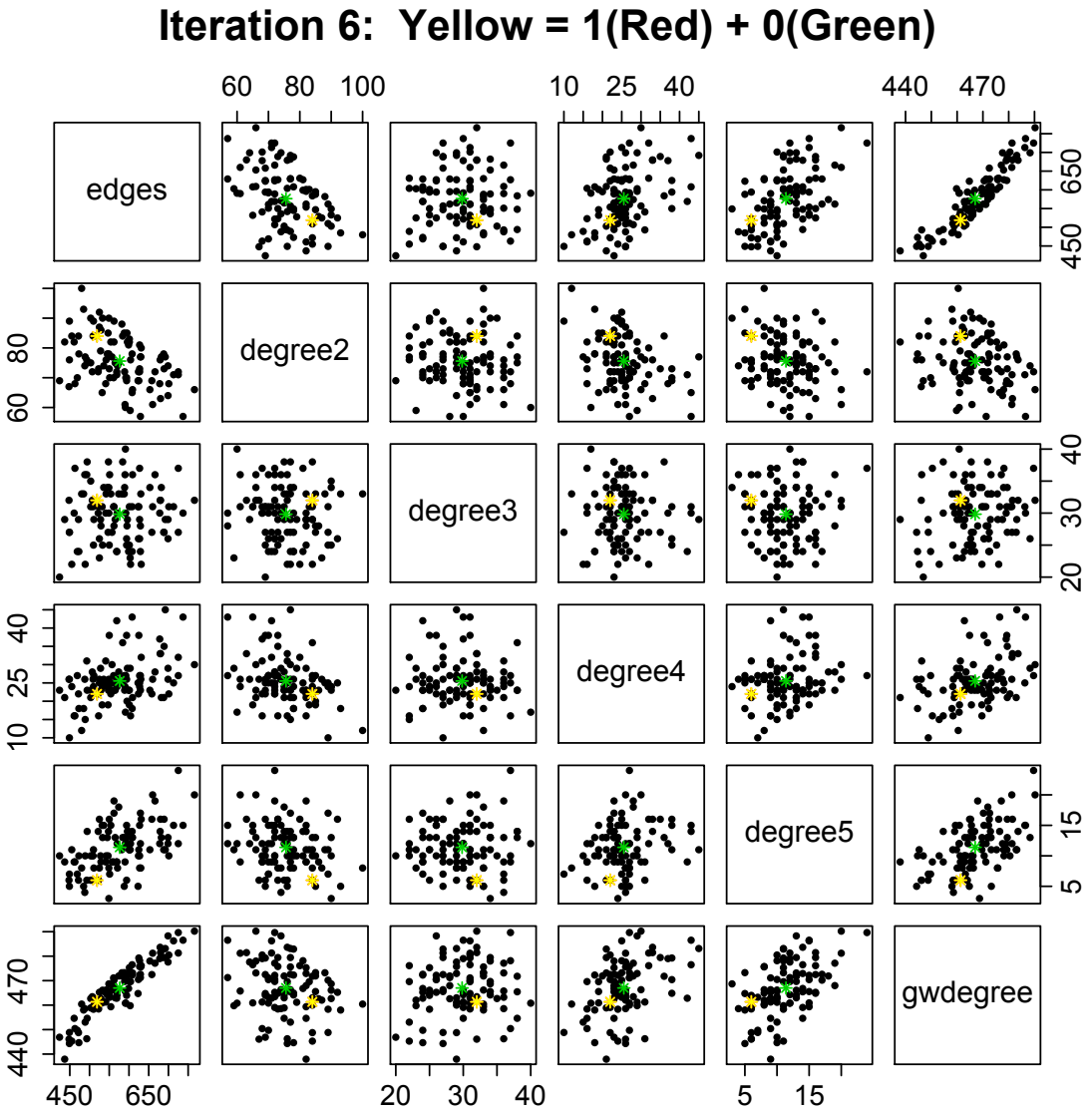
A sample from a model with a better θ_0



A sample from a model with a better θ_0

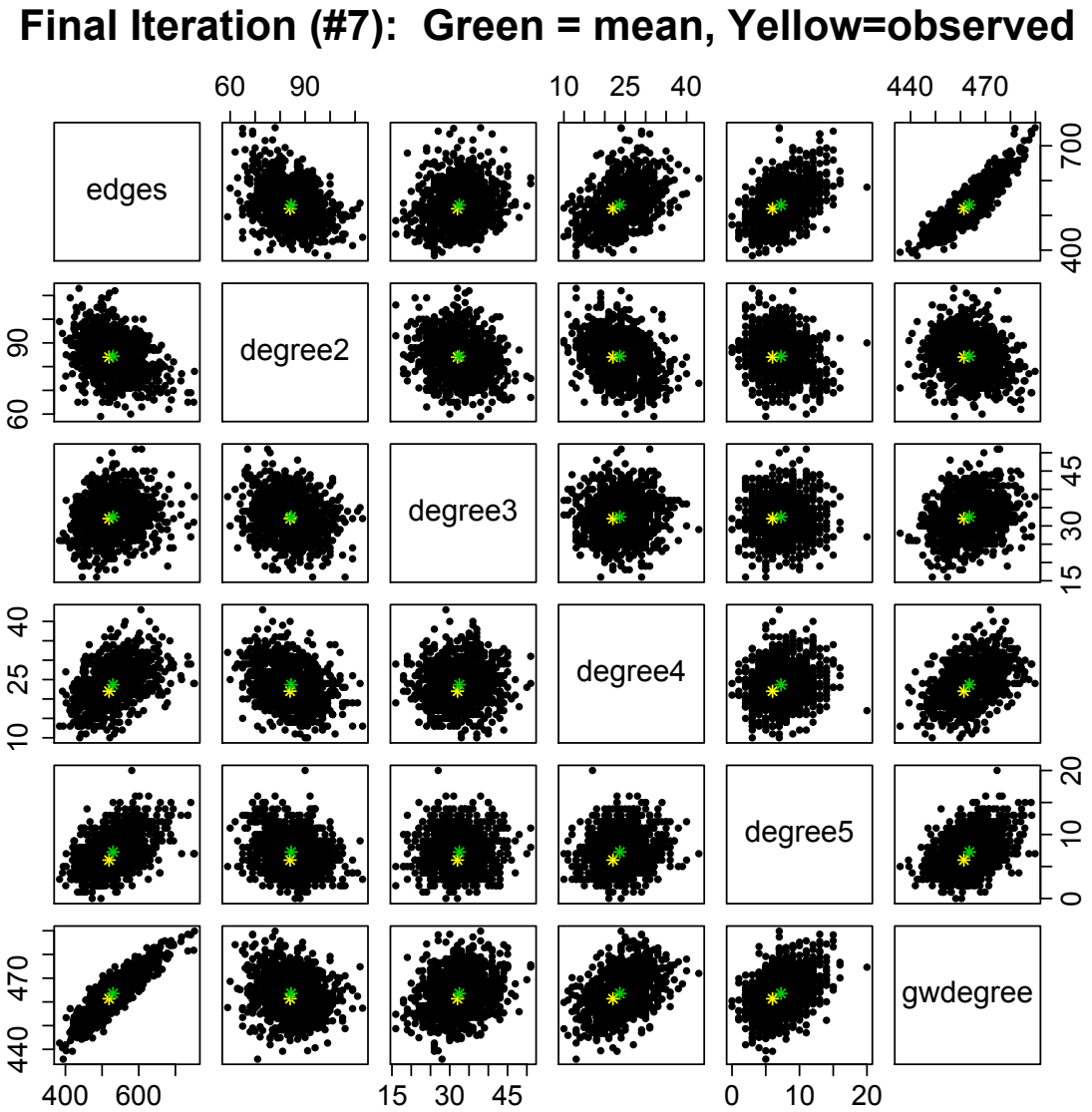


A sample from a model with a better θ_0



- Finally, we don't need to pretend; the true $g(y^{\text{obs}})$ is actually interior to the convex hull of sampled points. . .

A sample from a model with a better θ_0



- . . . so now we can take a larger sample and get a better final estimate of θ.

Finally, an MLE

Original MPLE:

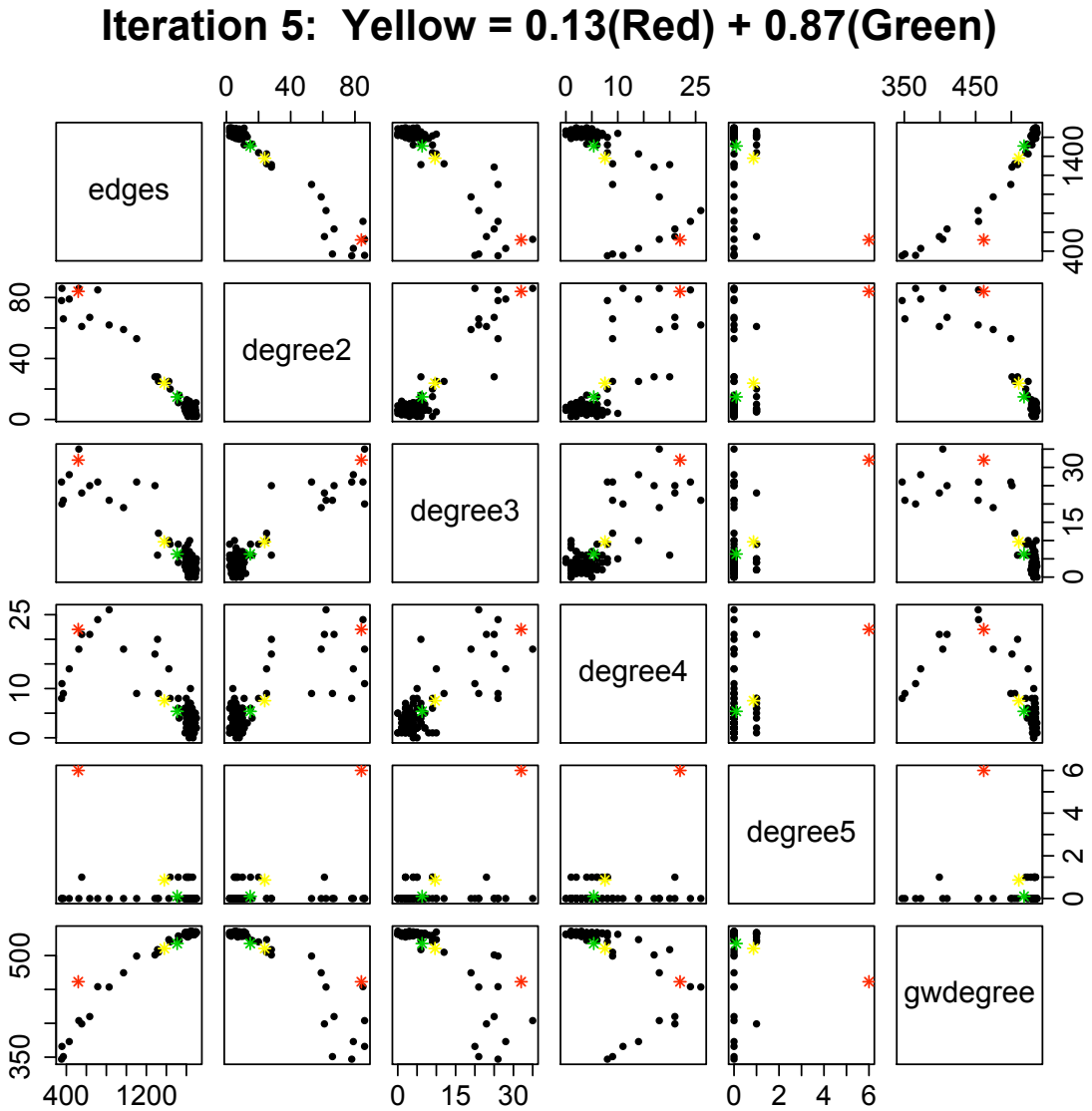
edges	degree2	degree3	degree4	degree5	gwdegree
-5.35	-2.58	-3.06	-2.39	-1.85	8.13

Final (approximated) MLE:

edges	degree2	degree3	degree4	degree5	gwdegree
-5.06	-1.45	-2.35	-2.28	-2.91	1.77

We know this is close to the true MLE $\hat{\theta}$ because the true MLE uniquely gives $E_{\hat{\theta}} g(Y) = g(y^{\text{obs}})$.

Failure of first approximation in this example



- A steplength of 0.01 is too small in this case.

Conclusions

- MPLE looks increasingly dangerous; it can mask problems when they exist and miss badly when they don't
- Naive MCMC MLE may not perform well even in very simple problems, but it may be modified. Here, we had success in a hard problem using two ideas:
  (a) Partial stepping toward the MLE in mean-value parameter space;
  (b) A log-normal approximation to the normalizing constant.
- By making MLE more and more automatic, we hope that scientists will be able to focus on modeling, not programming.

Outline

Cited References

- Alon, U. (2007, *Nature Reviews Genetics*), Network Motifs: Theory and Experimental Approaches.
- Barndorff-Nielsen, O. (1978). "Information and exponential families in statistical theory," New York: Wiley series in probability and mathematical statistics.
- Brown, L. D. (1986). "Fundamentals of statistical Exponential Families," Hayward: Institute of Mathematical Statistics.
- Saul ZM and Filkov V (2007, *Bioinformatics*), Exploring biological network structure using exponential random graph models.
- Salgado H et al. (2001, *Nucleic Acids Res.*), Regulondb (version 3.2): Transcriptional regulation and operon organization in Escherichia Coli k-12.
- van Duijn MAJ, Gile K, and Handcock MS (2009, *Social Networks*), A Framework for the Comparison of Maximum Pseudo-Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models.

Thank You!