Maximum Likelihood Estimation in Loglinear Models

Alessandro Rinaldo
Carnegie Mellon University
joint work with Steve Fienberg

April 18, 2012
Workshop on Graphical Models:
Mathematics, Statistics and Computer Sciences

Fields Institute

- Frequentist inference (estimation, goodness-of-fit testing, model selection) in log-linear models relies on the maximum likelihood estimator (MLE).

- Frequentist inference (estimation, goodness-of-fit testing, model selection) in log-linear models relies on the maximum likelihood estimator (MLE).

- The MLE may not exist due to sampling zeros.

- Frequentist inference (estimation, goodness-of-fit testing, model selection) in log-linear models relies on the maximum likelihood estimator (MLE).

- The MLE may not exist due to sampling zeros.

- Nonexistence of the MLE largely ignored in practice. Important issue, particular in large and sparse tables.

Motivating Pathological Example

- Consider a 2^3 table and the model [12][13][23]

Motivating Pathological Example

- Zero margin: MLE does not exist!

0	1
0	4

2	3
2	2

Motivating Pathological Example

- Haberman's example (1974). Positive margins and nonexistent MLE.

0	1
2	4

2	3
2	0

- `loglin` routine in `R`

```
Warning message:  Algorithm did not converge ...
         stop("This should not happen")
```

- `glm` routine in `R`

```
fitted rates numerically 0
```

Motivating Pathological Example

- Haberman's example (1974). Positive margins and nonexistent MLE.

0	1		2	3
2	4		2	0

- `loglin` routine in R

```
Warning message:  Algorithm did not converge ...
stop("This should not happen")
```

- `glm` routine in R

```
fitted rates numerically 0
```

- `PROC CATMOD` routine in SAS

```
If you want zero frequencies that PROC CATMOD would normally treat
as structural zeros to be interpreted as sampling zeros, simply
insert a one-line statement into the data step that changes each
zero to a very small number (such as 1E-20).
```

Motivating Pathological Example

- Haberman's example (1974). Positive margins and nonexistent MLE.

0	1
2	4

2	3
2	0

- loglin routine in R

```
Warning message:  Algorithm did not converge ...
stop("This should not happen")
```

- glm routine in R

```
fitted rates numerically 0
```

- PROC CATMOD routine in SAS

```
If you want zero frequencies that PROC CATMOD would normally treat
as structural zeros to be interpreted as sampling zeros, simply
insert a one-line statement into the data step that changes each
zero to a very small number (such as 1E-20).
```

- Always perfect fit to the data. The p-value is always 1.

Outline

- Background on log-linear models and exponential families.
- Existence of the MLE.
- Parameter estimability under a nonexistent MLE.
- Computation of extended MLE.

Log-Linear Models

- Consider the exponential family $\{P_\theta, \theta \in \mathbb{R}^{\mathcal{I}}\}$ over a finite set of *cells* $\mathcal{I}$:

$$P_\theta(\{i\}) = \exp\{(\theta, a_i) - \phi(\theta)\}, \quad \theta \in \mathbb{R}^d, i \in \mathcal{I},$$

with $a_i \in \mathbb{N}^d \setminus \{0\}$ and $\phi(\theta) = \log\left(\sum_i \exp\{(\theta, a_i)\}\right)$.
Model specified by a $|\mathcal{I}| \times d$ design matrix A, whose i-th row is $a_i^\top$.

Log-Linear Models

- Consider the exponential family $\{P_\theta, \theta \in \mathbb{R}^{\mathcal{I}}\}$ over a finite set of *cells* $\mathcal{I}$:

  $$P_\theta(\{i\}) = \exp\left\{(\theta, a_i) - \phi(\theta)\right\}, \quad \theta \in \mathbb{R}^d, i \in \mathcal{I},$$

  with $a_i \in \mathbb{N}^d \setminus \{0\}$ and $\phi(\theta) = \log\left(\sum_i \exp\{(\theta, a_i)\}\right)$.
  Model specified by a $|\mathcal{I}| \times d$ design matrix A, whose i-th row is $a_i^\top$.

- Observe a number N (possibly random) of cells $\{L_1, \ldots, L_N\}$, with $L_j \in \mathcal{I}$, all j.
  The corresponding contingency table is the random vector $n \in \mathbb{N}^{\mathcal{I}}$

  $$n(i) = |\{j : L_j = i\}|, \quad i \in \mathcal{I}.$$

Log-Linear Models

- Log-linear model analysis (see, e.g. Haberman, 1974 and Bishop et al., 2007) is concerned with modeling the distribution of n by assuming that
  - $m := \mathbb{E}(n) > 0$,
  - $\mu := \log(m) \in \mathcal{M} \subset \mathbb{R}^{\mathcal{I}}$, where $\mathcal{M} = \mathcal{R}(\mathrm{A})$ is the log-linear subspace.

- Sampling constraints: let $\mathcal{N} \subset \mathcal{M}$ be a linear subspace of $\mathcal{M}$ of dimension $m < d$: sampling subspace.

  Conditional Poisson sampling: $\{n(i), i \in \mathcal{I}\}$ have the conditional distribution of $|\mathcal{I}|$ independent Poisson random variables with means $\{\exp(\mu(i)), i \in \mathcal{I}\}$ given that $\Pi_{\mathcal{N}} n = c$ for some known $c \in \mathbb{R}^{\mathcal{I}}$.

Conditional Poisson Sampling Schemes: Examples

- Poisson Likelihood: $\mathcal{N} = \{0\}$. The log-likelihood is

$$(n, \mu) - \sum_i \exp(\mu(i)) - \sum_i \log\left(n(i)\right)!, \quad \mu \in \mathcal{M}.$$

Conditional Poisson Sampling Schemes: Examples

- Poisson Likelihood: $\mathcal{N} = \{0\}$. The log-likelihood is

$$(n, \mu) - \sum_i \exp(\mu(i)) - \sum_i \log(n(i))!, \quad \mu \in \mathcal{M}.$$

- Product Multinomial Likelihood: $\mathcal{N} = \mathrm{span}(\chi_1, \ldots, \chi_m)$, where the χ_j's are the indicator functions of a partition of $\mathcal{I}$. The sampling constrains are $(n, \chi_j) = N_j > 0$ for all j.

Conditional Poisson Sampling Schemes: Examples

- Poisson Likelihood: $\mathcal{N} = \{0\}$. The log-likelihood is

$$(n, \mu) - \sum_i \exp(\mu(i)) - \sum_i \log\,(n(i))!, \quad \mu \in \mathcal{M}.$$

- Product Multinomial Likelihood: $\mathcal{N} = \mathrm{span}(\chi_1, \ldots, \chi_m)$, where the χ_j's are the indicator functions of a partition of $\mathcal{I}$. The sampling constrains are $(n, \chi_j) = N_j > 0$ for all j.
  The log-likelihood is

$$\sum_{j=1}^{m} \left(\sum_{i \in \mathcal{B}_j} n(i) \log \frac{m(i)}{(m, \chi_j)} + \log N_j! - \sum_{i \in \mathcal{B}_j} \log n(i)! \right), \quad \mu \in \mathcal{M},$$

  well defined only on

$$\{\mu \in \mathcal{M} \colon (\chi_j, \exp(\mu)) = N_j, j = 1, \ldots, r\} \subset \mathcal{M}.$$

Conditional Poisson Sampling Schemes: Examples

- Poisson Likelihood: $\mathcal{N} = \{0\}$. The log-likelihood is

$$(n, \mu) - \sum_i \exp(\mu(i)) - \sum_i \log\,(n(i))!, \quad \mu \in \mathcal{M}.$$

- Product Multinomial Likelihood: $\mathcal{N} = \mathrm{span}(\chi_1, \ldots, \chi_m)$, where the χ_j's are the indicator functions of a partition of $\mathcal{I}$. The sampling constrains are $(n, \chi_j) = N_j > 0$ for all j.
  The log-likelihood is

$$\sum_{j=1}^{m} \left(\sum_{i \in \mathcal{B}_j} n(i) \log \frac{m(i)}{(m, \chi_j)} + \log N_j! - \sum_{i \in \mathcal{B}_j} \log n(i)! \right), \quad \mu \in \mathcal{M},$$

  well defined only on

$$\{\mu \in \mathcal{M} \colon (\chi_j, \exp(\mu)) = N_j, j = 1, \ldots, r\} \subset \mathcal{M}.$$

- Poisson-Multinomial likelihood (Lang, 2005): a combination of the two.

Example: Hierarchical Log-Linear Models

- Let $X_1, \ldots, X_K$ be discrete random variables, each X_k supported on finite set $\mathcal{I}_k$ of labels. Then $\mathcal{I} = \times_{k=1}^{K} \mathcal{I}_k$.
  A hierarchical log-linear model Δ is a simplicial complex: class of subsets of $\{1, \ldots, K\}$ such that $S \in \Delta$ and $T \subset S$ implies $T \in \Delta$).
  Graphical models are special cases.
- There log-linear subspaces are of ANOVA-type and there are canonical ways of constructing A (see Lauritzen, 1996, Appendix B).

Example: Hierarchical Log-Linear Models

- Let $X_1, \ldots, X_K$ be discrete random variables, each X_k supported on finite set $\mathcal{I}_k$ of labels. Then $\mathcal{I} = \times_{k=1}^{K} \mathcal{I}_k$.
  A hierarchical log-linear model Δ is a simplicial complex: class of subsets of $\{1, \ldots, K\}$ such that $S \in \Delta$ and $T \subset S$ implies $T \in \Delta$).
  Graphical models are special cases.
- There log-linear subspaces are of ANOVA-type and there are canonical ways of constructing A (see Lauritzen, 1996, Appendix B).

- Inferential tasks: estimation, goodness-of-fit testing and model selection.

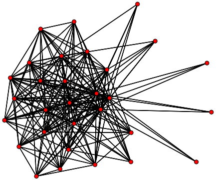
Example: Hierarchical Log-Linear Models

Mildew fungus example: 2^6 sparse table. Source: Edwards (2000).

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

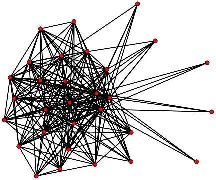
Example: Network Models – the β-Model

- Let $\mathcal{G}_v$ be the set of simple graphs on v nodes.



Example: Network Models – the β-Model

- Let $\mathcal{G}_v$ be the set of simple graphs on v nodes.

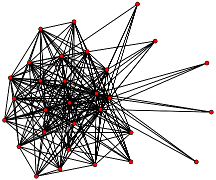


- β-Model: edges are independent and occurs with probabilities

$$\frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}}, \quad i \neq j, \quad \beta = (\beta_1, \ldots, \beta_v) \in \mathbb{R}^v.$$

Example: Network Models – the β-Model

- Let $\mathcal{G}_v$ be the set of simple graphs on v nodes.



- β-Model: edges are independent and occurs with probabilities

$$\frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}, \quad i \neq j, \quad \beta = (\beta_1, \ldots, \beta_v) \in \mathbb{R}^v.$$

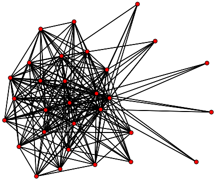
The probability of a graph $x \in \mathcal{G}_n$ is

$$\exp\left\{ \sum_{i=1}^{v} d_i \beta_i - \psi(\beta) \right\}, \quad \beta \in \mathbb{R}^v,$$

where $d(x) = d = (d_1, \ldots, d_v)$ is the (ordered) degree sequence of x.

Example: Network Models – the β-Model

- Let $\mathcal{G}_v$ be the set of simple graphs on v nodes.



- β-Model: edges are independent and occurs with probabilities

$$\frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}}, \quad i \neq j, \quad \beta = (\beta_1, \ldots, \beta_v) \in \mathbb{R}^v.$$

The probability of a graph $x \in \mathcal{G}_n$ is

$$\exp\left\{\sum_{i=1}^{v} d_i\beta_i - \psi(\beta)\right\}, \quad \beta \in \mathbb{R}^v,$$

where $d(x) = d = (d_1, \ldots, d_v)$ is the (ordered) degree sequence of x.

- It is a log-linear model under product multinomial sampling.

Exponential Family Representation

- It is more convenient to represent the log-linear model likelihood in exponential form, with densities

$$p_\theta(n) = \exp\left\{(\mathrm{A}^\top n, \theta) - \psi(\theta)\right\}\nu(n), \quad \theta \in \mathbb{R}^d,$$

where $n \in S(\mathcal{N}, c) := \{x \in \mathbb{N}^{\mathcal{I}} \colon \Pi_{\mathcal{N}} x = c\}$ and the base measure is

$$\nu(x) = 1_{x \in S(\mathcal{N},c)} \prod_{i \in \mathcal{I}} \frac{1}{x(i)!}, \quad x \in \mathbb{N}^{\mathcal{I}}.$$

Exponential Family Representation

- It is more convenient to represent the log-linear model likelihood in exponential form, with densities

$$p_\theta(n) = \exp\left\{ (\mathrm{A}^\top n, \theta) - \psi(\theta) \right\} \nu(n), \quad \theta \in \mathbb{R}^d,$$

  where $n \in S(\mathcal{N}, c) := \{x \in \mathbb{N}^{\mathcal{I}} \colon \Pi_{\mathcal{N}} x = c\}$ and the base measure is

$$\nu(x) = 1_{x \in S(\mathcal{N}, c)} \prod_{i \in \mathcal{I}} \frac{1}{x(i)!}, \quad x \in \mathbb{N}^{\mathcal{I}}.$$

- It is a family of order $d - m$, where $m = \dim(\mathcal{N}) > 0$. Minimality: replace A with full-rank A' such that $\mathcal{R}(\mathrm{A}') = \mathcal{M} \ominus \mathcal{N} := \mathcal{M} \cap \mathcal{N}^c$.

Exponential Family Representation

- It is more convenient to represent the log-linear model likelihood in exponential form, with densities

$$p_\theta(n) = \exp\left\{(\mathrm{A}^\top n, \theta) - \psi(\theta)\right\}\nu(n), \quad \theta \in \mathbb{R}^d,$$

where $n \in S(\mathcal{N}, c) := \{x \in \mathbb{N}^{\mathcal{I}} \colon \Pi_{\mathcal{N}} x = c\}$ and the base measure is

$$\nu(x) = 1_{x \in S(\mathcal{N},c)} \prod_{i \in \mathcal{I}} \frac{1}{x(i)!}, \quad x \in \mathbb{N}^{\mathcal{I}}.$$

- It is a family of order $d - m$, where $m = \dim(\mathcal{N}) > 0$. Minimality: replace A with full-rank A' such that $\mathcal{R}(\mathrm{A}') = \mathcal{M} \ominus \mathcal{N} := \mathcal{M} \cap \mathcal{N}^c$.
- Better log-likelihood model parametrization for product-multinomial sampling:

$$(n, \beta) - \sum_{j=1}^{m} N_j \log(\exp^\beta, \chi_j) - \sum_{i \in \mathcal{I}} \log n(i)!, \quad \beta \in \mathcal{M} \ominus \mathcal{N}.$$

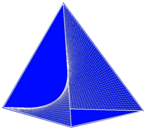
Three views of Log-Linear Models

1. The exponential family natural parametrization: $\mathbb{R}^{d-m}$.
2. The log-linear model parametrization: $\mathcal{M} \ominus \mathcal{N} \subset \mathbb{R}^{\mathcal{I}}$.

Three views of Log-Linear Models

1. The exponential family natural parametrization: $\mathbb{R}^{d-m}$.
2. The log-linear model parametrization: $\mathcal{M} \ominus \mathcal{N} \subset \mathbb{R}^{\mathcal{I}}$.
3. Connection with Algebraic Geometry (see, e.g., Drton et al. 2008).
   Let $M = \{\exp^{\mu}, \mu \in \mathcal{M}\}$: intersection of a real toric variety in $\mathbb{R}^{\mathcal{I}}$ with the positive orthant.
   The model is parametrized by

$$V = M \cap \{x \in \mathbb{R}^{\mathcal{I}} : \Pi_{\mathcal{N}} x = c\}.$$



Interpretable parametrization: (conditional) expected cell counts.

Maximum Likelihood Estimation

In log-linear models inference relies on MLE:

$$\widehat{\theta} = \mathrm{argsup}_{\theta \in \mathbb{R}^{d-m}} p_\theta(n)$$

The MLEs $\widehat{\mu} \in \mathcal{M} \ominus \mathcal{N}$ and $\widehat{m} \in V$ are similarly defined.

Maximum Likelihood Estimation

In log-linear models inference relies on MLE:

$$\widehat{\theta} = \mathrm{argsup}_{\theta \in \mathbb{R}^{d-m}} p_{\theta}(n)$$

The MLEs $\widehat{\mu} \in \mathcal{M} \ominus \mathcal{N}$ and $\widehat{m} \in V$ are similarly defined.

- If the supremum is not achieved the MLE does not exist. Instead supremum is realized in the limit
  - $\{\theta_n\} \subset \mathbb{R}^{d-m}$: $\|\theta_n\| \to \infty$
  - $\{\mu_n\} \subset \mathcal{M} \ominus \mathcal{N}$: $\|\mu_n\| \to \infty$
  - $\{m_n\} \subset V$: $m_n(i) \to 0$, for some i

Maximum Likelihood Estimation

In log-linear models inference relies on MLE:

$$\widehat{\theta} = \mathrm{argsup}_{\theta \in \mathbb{R}^{d-m}} p_\theta(n)$$

The MLEs $\widehat{\mu} \in \mathcal{M} \ominus \mathcal{N}$ and $\widehat{m} \in V$ are similarly defined.

- If the supremum is not achieved the MLE does not exist. Instead supremum is realized in the limit
  - $\{\theta_n\} \subset \mathbb{R}^{d-m}$: $\|\theta_n\| \to \infty$
  - $\{\mu_n\} \subset \mathcal{M} \ominus \mathcal{N}$: $\|\mu_n\| \to \infty$
  - $\{m_n\} \subset V$: $m_n(i) \to 0$, for some i
- Far from being a just numerical issue. Existent MLE needed to
  - get correct asymptotic approximations to various goodness-of-fit testing statistics in regular and double-asymptotic setting;
  - obtain standard errors for the model parameters;
  - obtain proper posteriors under improper priors and adequate Bayesian inference;
  - carry out conditional inference.

Maximum Likelihood Estimation

In log-linear models inference relies on MLE:

$$\widehat{\theta} = \text{argsup}_{\theta \in \mathbb{R}^{d-m}} p_\theta(n)$$

The MLEs $\widehat{\mu} \in \mathcal{M} \ominus \mathcal{N}$ and $\widehat{m} \in V$ are similarly defined.

- If the supremum is not achieved the MLE does not exist. Instead supremum is realized in the limit
  - $\{\theta_n\} \subset \mathbb{R}^{d-m}$: $\|\theta_n\| \to \infty$
  - $\{\mu_n\} \subset \mathcal{M} \ominus \mathcal{N}$: $\|\mu_n\| \to \infty$
  - $\{m_n\} \subset V$: $m_n(i) \to 0$, for some i
- Far from being a just numerical issue. Existent MLE needed to
  - get correct asymptotic approximations to various goodness-of-fit testing statistics in regular and double-asymptotic setting;
  - obtain standard errors for the model parameters;
  - obtain proper posteriors under improper priors and adequate Bayesian inference;
  - carry out conditional inference.
- Nonexistence of the MLE leads to issues of estimability and assessment of the model complexity.

Maximum Likelihood Estimation

- Well known issue in theory: Haberman (1974), Aickin (1979), Glonek et al. (1988), Verbeek (1992) Glonek, Lauritzen (1996).

Maximum Likelihood Estimation

- Well known issue in theory: Haberman (1974), Aickin (1979), Glonek et al. (1988), Verbeek (1992) Glonek, Lauritzen (1996).
- Haberman's pathological example (1974).

[12][13][23]

0	

	0

Maximum Likelihood Estimation

- Well known issue in theory: Haberman (1974), Aickin (1979), Glonek et al. (1988), Verbeek (1992) Glonek, Lauritzen (1996).
- Haberman's pathological example (1974).

[12][13][23]

0	

	0

- In 2^K table with positive entries set 2 cells to zero at random. Under the model of no-3-way interactions, a zero margin will occur with probability

$$\frac{k}{2^k - 1} \approx 0, \text{ for large } K,$$

and a nonexistent MLE leaving all margins positive with probability

$$\frac{2^{k-1} - k}{2^k - 1} \approx \frac{1}{2}, \text{ for large } K.$$

Maximum Likelihood Estimation: Goals

1. Characterize patterns of sampling zeros leading to the nonexistence of the MLE.
2. Statistical consequence of a nonexistent MLE.
3. Algorithms.

Basics of Discrete Exponential Families

See Barndorff-Nielsen (1974), Brown (1986), Jensen (1989), Letác (1992), Csiszár and Matúš (2001,2003,2005,2008).

Basics of Discrete Exponential Families

See Barndorff-Nielsen (1974), Brown (1986), Jensen (1989), Letác (1992), Csiszár and Matúš (2001,2003,2005,2008).

- The distribution of $T = \mathrm{A}^\top n$ belongs to the family $\mathcal{E} = \{P_\theta, \theta \in \Theta\}$, with

$$P_\theta(t) = \exp\{\langle\theta, t\rangle - \psi(\theta)\}\,\mu(t), \quad \theta \in \Theta = \mathbb{R}^{d-m}.$$

  whose support $\mathcal{T} \subset \mathbb{R}^d$ and base measure μ depend on the sampling scheme.

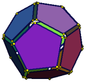
Basics of Discrete Exponential Families

See Barndorff-Nielsen (1974), Brown (1986), Jensen (1989), Letác (1992), Csiszár and Matúš (2001,2003,2005,2008).

- The distribution of $T = \mathrm{A}^\top n$ belongs to the family $\mathcal{E} = \{P_\theta, \theta \in \Theta\}$, with
$$P_\theta(t) = \exp\{\langle\theta, t\rangle - \psi(\theta)\}\,\mu(t), \quad \theta \in \Theta = \mathbb{R}^{d-m}.$$
whose support $\mathcal{T} \subset \mathbb{R}^d$ and base measure μ depend on the sampling scheme.
- The set $\mathrm{P} = \mathrm{convhull}(\mathcal{T})$ is called the convex support of $\mathcal{E}$.

Basics of Discrete Exponential Families

See Barndorff-Nielsen (1974), Brown (1986), Jensen (1989), Letác (1992), Csiszár and Matúš (2001,2003,2005,2008).

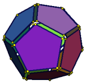
- The distribution of $T = \mathrm{A}^\top n$ belongs to the family $\mathcal{E} = \{P_\theta, \theta \in \Theta\}$, with
$$P_\theta(t) = \exp\{\langle\theta, t\rangle - \psi(\theta)\}\,\mu(t), \quad \theta \in \Theta = \mathbb{R}^{d-m}.$$
whose support $\mathcal{T} \subset \mathbb{R}^d$ and base measure μ depend on the sampling scheme.
- The set $\mathrm{P} = \mathrm{convhull}(\mathcal{T})$ is called the convex support of $\mathcal{E}$.
    - It is a polyhedron. 
    - $\mathrm{int}(\mathrm{P}) = \{\mathbb{E}_\theta[T], \theta \in \Theta\}$: mean value space.
    - $\mathrm{int}(\mathrm{P})$ and Θ are homeomorphic: mean value parametrization.

Basics of Discrete Exponential Families

See Barndorff-Nielsen (1974), Brown (1986), Jensen (1989), Letác (1992), Csiszár and Matúš (2001,2003,2005,2008).

- The distribution of $T = \mathrm{A}^\top n$ belongs to the family $\mathcal{E} = \{P_\theta, \theta \in \Theta\}$, with

$$P_\theta(t) = \exp\left\{\langle \theta, t\rangle - \psi(\theta)\right\}\mu(t), \quad \theta \in \Theta = \mathbb{R}^{d-m}.$$

  whose support $\mathcal{T} \subset \mathbb{R}^d$ and base measure μ depend on the sampling scheme.
- The set $\mathrm{P} = \mathrm{convhull}(\mathcal{T})$ is called the convex support of $\mathcal{E}$.
  - It is a polyhedron. 
  - $\mathrm{int}(\mathrm{P}) = \{\mathbb{E}_\theta[T], \theta \in \Theta\}$: mean value space.
  - $\mathrm{int}(\mathrm{P})$ and Θ are homeomorphic: mean value parametrization.

Existence of the MLE

The MLE exists if and only if $t \in \mathrm{int}(\mathrm{P})$.

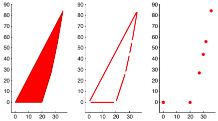
Extended Exponential Families: Geometric Construction

- For every face F of P, construct the exponential family of distributions $\mathcal{E}_F$ for the sample points in F with convex support F. Note that $\mathcal{E}_F$ depends on $\dim(F) < d$ parameters only.

Extended Exponential Families: Geometric Construction

- For every face F of P, construct the exponential family of distributions $\mathcal{E}_F$ for the sample points in F with convex support F. Note that $\mathcal{E}_F$ depends on $\dim(F) < d$ parameters only.

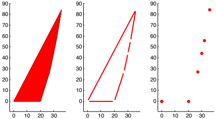
- The extended exponential family is $\overline{\mathcal{E}} = \mathcal{E} \cup \{\mathcal{E}_F : F \text{ is a face of } \mathrm{P}\}$.



Extended Exponential Families: Geometric Construction

- For every face F of P, construct the exponential family of distributions $\mathcal{E}_F$ for the sample points in F with convex support F. Note that $\mathcal{E}_F$ depends on $\dim(F) < d$ parameters only.

- The extended exponential family is $\overline{\mathcal{E}} = \mathcal{E} \cup \{\mathcal{E}_F : F \text{ is a face of } \mathrm{P}\}$.

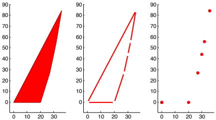


- Within $\overline{\mathcal{E}}$, the MLE always exists.

Extended Exponential Families: Geometric Construction

- For every face F of P, construct the exponential family of distributions $\mathcal{E}_F$ for the sample points in F with convex support F. Note that $\mathcal{E}_F$ depends on $\dim(F) < d$ parameters only.

- The extended exponential family is $\overline{\mathcal{E}} = \mathcal{E} \cup \{\mathcal{E}_F : F \text{ is a face of } \mathrm{P}\}$.

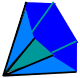
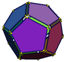
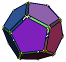


- Within $\overline{\mathcal{E}}$, the MLE always exists.

Extended Exponential Family

The extended exponential family is the closure of the original family. Geometrically, this corresponds to taking the closure of the mean value space, i.e. including the boundary of P.

Convex Supports (Mean Value Parametrization)

- Poisson sampling scheme: marginal cone $C_A = \mathrm{cone}(A)$ (Eriksson et al., 2006) 
- Multinomial sampling scheme: marginal polytope $\mathrm{convhull}(A)$ 
- Product multinomial sampling scheme: Minkowksi sum of convex polytopes  + ... + 
- Poisson-multinomial: Miknowksi sum of polyhedral cone and convex polytopes  + ... +  + 

Convex Supports (Mean Value Parametrization)

- $\mathrm{int}(\mathrm{P})$ homeomorphic to V, with homeomorpism given by

$$x \in V \mapsto \mathrm{A}^{\top} x \in P,$$

known as the moment map.

Convex Supports (Mean Value Parametrization)

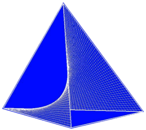
- $\mathrm{int}(\mathrm{P})$ homeomorphic to V, with homeomorpism given by

$$x \in V \mapsto \mathrm{A}^\top x \in P,$$

  known as the moment map.

- Homeomorphism extended to the boundaries.

$\mathrm{cl}(V)$ is also a mean value parametrization of $\overline{\mathcal{E}}$.



Maximum Likelihood Estimation: Existence

Existence of the MLE (assume minimality)

The MLE of θ (or of μ or of m) exists (and is unique) if and only if $\mathrm{A}^\top n \in \mathrm{ri}(\mathrm{C_A})$ and satisfy the moment equations

$$\nabla\psi(\widehat{\theta}) = \mathrm{A}^\top n$$

or, equivalently,

$$\widehat{m} = \widehat{m}(\widehat{\theta}) = V \cap \{x \geqslant 0\colon \mathrm{A}^\top x = \mathrm{A}^\top n\}.$$

Maximum Likelihood Estimation: Existence

Existence of the MLE (assume minimality)

The MLE of θ (or of μ or of m) exists (and is unique) if and only if $\mathrm{A}^\top n \in \mathrm{ri}(\mathrm{C}_\mathrm{A})$ and satisfy the moment equations

$$\nabla \psi(\widehat{\theta}) = \mathrm{A}^\top n$$

or, equivalently,

$$\widehat{m} = \widehat{m}(\widehat{\theta}) = V \cap \{x \geqslant 0 \colon \mathrm{A}^\top x = \mathrm{A}^\top n\}.$$

- Application of standard theory of exponential families. Haberman (1974) first to derive it.
- Results apply to more general conditional Poisson sampling under additional condition.

Maximum Likelihood Estimation: Existence

Need more explicit result to characterize problematic zero entries in the table.

Maximum Likelihood Estimation: Existence

Need more explicit result to characterize problematic zero entries in the table.

- Facial sets (Geiger et al., 2006). A set $\mathcal{F} \subseteq \mathcal{I}$ for which, for some $c \in \mathbb{R}^d$,

$$\begin{array}{ll} (a_i, c) = 0 & i \in \mathcal{F} \\ (a_i, c) < 0 & i \notin \mathcal{F}, \end{array}$$

where a_i is the i-th row of A is called a facial set of C_{A}.

Maximum Likelihood Estimation: Existence

Need more explicit result to characterize problematic zero entries in the table.

- Facial sets (Geiger et al., 2006). A set $\mathcal{F} \subseteq \mathcal{I}$ for which, for some $c \in \mathbb{R}^d$,

$$\begin{array}{ll} (a_i, c) = 0 & i \in \mathcal{F} \\ (a_i, c) < 0 & i \notin \mathcal{F}, \end{array}$$

  where a_i is the i-th row of A is called a facial set of $\mathrm{C_A}$.
- Facial sets captures combinatorial structure of $\mathrm{C_A}$ and of $\mathrm{cl}(V)$. The following are equivalent
  - $t \in \mathrm{ri}(F)$
  - $t \in \mathrm{cone}\left(\{a_i, i \in \mathcal{F}\}\right)$
  - $t = \mathrm{A}^\top m$, for one $m \in \mathrm{cl}(V)$ with $\mathrm{supp}(m) = \mathcal{F}$.

Maximum Likelihood Estimation: Existence

Need more explicit result to characterize problematic zero entries in the table.

- Facial sets (Geiger et al., 2006). A set $\mathcal{F} \subseteq \mathcal{I}$ for which, for some $c \in \mathbb{R}^d$,

$$\begin{array}{ll} (a_i, c) = 0 & i \in \mathcal{F} \\ (a_i, c) < 0 & i \notin \mathcal{F}, \end{array}$$

  where a_i is the i-th row of A is called a facial set of $\mathrm{C_A}$.
- Facial sets captures combinatorial structure of $\mathrm{C_A}$ and of $\mathrm{cl}(V)$. The following are equivalent
  - $t \in \mathrm{ri}(F)$
  - $t \in \mathrm{cone}\left(\{a_i, i \in \mathcal{F}\}\right)$
  - $t = \mathrm{A}^\top m$, for one $m \in \mathrm{cl}(V)$ with $\mathrm{supp}(m) = \mathcal{F}$.

Existence of the MLE

The MLE does not exists if and only if $\{i \colon n(i) = 0\} \supset \mathcal{F}^c$, for a facial set $\mathcal{F}$.

Examples: Likelihood Zeros – `polymake`

2^2 table and the model [12][13][23].

$\mathrm{C_A}$ has 16 facets, 12 of which correspond to null margins.

0	
0	

Examples: Likelihood Zeros – `polymake`

2^2 table and the model [12][13][23].
C_A has 16 facets, 12 of which correspond to null margins.

0	

	0

Examples: Likelihood Zeros – `polymake`

3^3 table and the model [12][13][23].
$\mathrm{C_A}$ has 207 facets, 27 of which correspond to null margins.

0	0	
0	0	

0	0	
0	0	

		0

Examples: Likelihood Zeros – `polymake`

3^3 table and the model [12][13][23].
$\mathrm{C_A}$ has 207 facets, 27 of which correspond to null margins.

0		
0	0	

	0	0
	0	

0		0
		0

Examples: Likelihood Zeros – `polymake`

$4 \times 3 \times 6$ table and the model [12][13][23].
$\mathrm{C_A}$ has 153,858 facets, 54 of which correspond to null margins.

0		
0		0
0		0

	0	0
		0
0		0

	0	
	0	0
	0	0

	0	
	0	0
	0	0

	0	0
		0
0		0

0	0	
	0	
0		

Examples: Likelihood Zeros – polymake

2^4 table and the non-graphical model [12][13][14][23][34].
C_{A} has 56 facets, 24 of which correspond to zero margins.

0	0
0	

0	

0	

	0

Examples: Likelihood Zeros – `polymake`

3^4 table and the 4-cycle model [12][14][23][34].
C_A has 1,116 facets, 16 of which correspond to zero margins.

0		0
		0

0	0	0
0	0	0
0	0	

0	0	0
	0	0
	0	

0		0
0	0	0
0	0	0

0		
0	0	0
0	0	

	0	0
	0	

0		0
		0
0	0	0

0		
0	0	

0	0	0
	0	0
0	0	0

Example: The β-Model

- For the β-model, the convex support is the polytope of degree sequences: $P_v \subset \mathbb{R}^v$.

Example: The β-Model

- For the β-model, the convex support is the polytope of degree sequences: $P_v \subset \mathbb{R}^v$.

  The MLE is not defined whenever $d_i = 0$ or $d_i = v - 1$ for some i. These are $2v$ (easy) cases in total; many more...

Example: The β-Model

- For the β-model, the convex support is the polytope of degree sequences: $P_v \subset \mathbb{R}^v$.

  The MLE is not defined whenever $d_i = 0$ or $d_i = v - 1$ for some i. These are $2v$ (easy) cases in total; many more...

- The combinatorial complexity of P_v is large.
  Example the f-vector of P_8 is (Stanley, 1991)

  $$(334982, 1726648, 3529344, 3679872, 2074660, 610288, 81144, 3322).$$

  The number of facets and of vertices of P_4, P_5, P_6 and P_7 are 22, 60, 224 and 882 and 46, 332, 2874 and 29874, respectively.

Example: The β-Model

- For the β-model, the convex support is the polytope of degree sequences: $P_v \subset \mathbb{R}^v$.

  The MLE is not defined whenever $d_i = 0$ or $d_i = v - 1$ for some i. These are $2v$ (easy) cases in total; many more...

- The combinatorial complexity of P_v is large.
  Example the f-vector of P_8 is (Stanley, 1991)

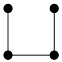
  $$(334982, 1726648, 3529344, 3679872, 2074660, 610288, 81144, 3322).$$

  The number of facets and of vertices of P_4, P_5, P_6 and P_7 are 22, 60, 224 and 882 and 46, 332, 2874 and 29874, respectively.

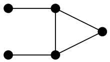
- Cayley Trick: Represent β-model as a log-linear model under product-multinomial constraints. Use a higher-dimensional marginal cone in $\mathbb{R}^{\binom{v}{2}+v}$ with smaller complexity.

Example: The β-Model

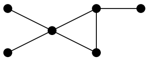
- When $v = 4$, there are 14 facial sets corresponding to the facets of P_4, 8 of which associated to a degree of 0 or 3.



- For $v = 5$, example of a facial set for which the degrees are bounded away from 0 and 4.



- For $v = 6$, example of a facial set for which the degrees are bounded away from 0 and 5.



Parameter Estimability

Assume the Poisson scheme and suppose that the MLE does not exist.

- $\mathrm{A}^\top n \in \mathrm{ri}(F)$, for some random face F of $\mathrm{C_A}$ of random dimension d_F.
- Maximum likelihood estimation well-defined in $\mathcal{E}_F$.

What can we do (estimate)?

Parameter Estimability

Assume the Poisson scheme and suppose that the MLE does not exist.

- $\mathrm{A}^\top n \in \mathrm{ri}(F)$, for some random face F of $\mathrm{C_A}$ of random dimension d_F.
- Maximum likelihood estimation well-defined in $\mathcal{E}_F$.

What can we do (estimate)?

- Let $\mathcal{L}_F$ to be the subspace generated by the normal cone to F. Equivalence relation on $\mathbb{R}^d$: $\theta_1 \overset{\mathcal{L}_F}{\sim} \theta_2$ if and only if $\theta_1 - \theta_2 \in \mathcal{L}_F$. Set $\theta_{\mathcal{L}_F}$ for the equivalence class containing θ and $\Theta_{\mathcal{L}_F} = \{\theta_{\mathcal{L}_F}, \theta \in \mathbb{R}^d\}$.
- Let $\mathcal{F}$ the facial set associated to F and let $\pi_{\mathcal{F}} : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{F}}$ be the coordinate projection:

$$x \mapsto \{x(i) : i \in \mathcal{F}\}.$$

Parameter Estimability

Parameter Estimability

(i) The family $\mathcal{E}_F$ is non-identifiable: any two points $\theta_1 \stackrel{\mathcal{L}_F}{\sim} \theta_2$ specify the same distribution.

(ii) The family $\mathcal{E}_F$ is parametrized by $\Theta_{\mathcal{L}_F}$, or, equivalently, by $\pi_{\mathcal{F}}(\mathcal{M})$ and is of order d_F.

(iii) The set $\Theta_{\mathcal{L}_F}$ is a d_F-dimensional dimensional vector space comprised of parallel affine subspaces of $\mathbb{R}^d$ of dimension $\dim(\mathcal{L}_F) = d - d_F$. It is isomorphic to $\pi_{\mathcal{F}}(\mathcal{M})$.

For product multinomial sampling schemes, replace

- $\mathcal{L}_F$ with $\mathcal{L}_F + \{\zeta \colon \mathrm{A}\zeta \in \mathcal{N}\}$;
- $\pi_{\mathcal{F}}(\mathcal{M})$ with $\pi_{\mathcal{F}}(\mathcal{M} \ominus \mathcal{N})$.

Parameter Estimability

- For the extended family $\mathcal{E}_F$ with corresponding facial set $\mathcal{F}$
  - $\mathcal{M} \cap (\mathcal{N} + \mathcal{L}_F)^c$ is estimable
  - $\pi_{\mathcal{F}}(V)$ is estimable

where $\mathcal{L}_F$, $\mathcal{F}$ and their dimension d_F are random.

Parameter Estimability

- For the extended family $\mathcal{E}_F$ with corresponding facial set $\mathcal{F}$
  - $\mathcal{M} \cap (\mathcal{N} + \mathcal{L}_F)^c$ is estimable
  - $\pi_{\mathcal{F}}(V)$ is estimable

  where $\mathcal{L}_F$, $\mathcal{F}$ and their dimension d_F are random.

- Extended MLEs:
  - Natural parametrization: Reparametrize $\mathcal{E}_F$ using a new design mantrix $\mathrm{A}_{\mathcal{F}}$ with $\mathcal{R}(\mathrm{A}_{\mathcal{F}}) = \mathcal{M} \cap (\mathcal{N} + \mathcal{L}_F)^c$. The extended MLE of the natural parameter is the MLE (which exists and is unique) of the corresponding family.
  - Mean value parametrization: The extended MLE is the unique point

$$\widehat{m} = \mathrm{bd}(V) \cap \{x \geqslant 0 \colon \mathrm{A}^\top x = \mathrm{A}^\top n\},$$

    where $\mathrm{supp}(\widehat{m}) = \mathcal{F}$.

- Correct model complexity: the adjusted number of degrees of freedom is

$$|\mathcal{F}| - d_F.$$

Parameter Estimability: Examples

2^3 table and the model [12][13][23]

0	

	0

$d_F = |\mathcal{F}| = 6$: saturated model for $\mathcal{E}_F$!

Parameter Estimability: Examples

3^3 table and the model [12][13][23]

0		
0	0	

	0	0
	0	

0		0
		0

$d_F = |\mathcal{F}| = 18$: saturated model for $\mathcal{E}_F$!

Parameter Estimability: Examples

3^3 table and the model [12][13][23]. The red zeros are not likelihood zeros.

0		
0		

		0
	0	0

0		**0**
0		

$d_F = 18$, $|\mathcal{F}| = 21$: the number of adjusted degrees of freedom is 3.

Parameter Estimability: Examples

3^3 table and the model [12][13][23]

	0	0
0		0
0	0	

0	0	
	0	0
0		0

0		0
0	0	
	0	0

The MLE exists!

Parameter Estimability: Mildew Fungus Example

A	B	C	D=1				D=2			
		E	1		2		1		2	
		F	1	2	1	2	1	2	1	2
1	1	1	0	0	0	0	3	0	1	0
		2	0	1	0	0	0	1	0	0
	2	1	1	0	1	0	7	1	4	0
		2	0	0	0	2	1	3	0	11
2	1	1	16	1	4	0	1	0	0	0
		2	1	4	1	4	0	0	0	1
	2	1	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0

Parameter Estimability: Mildew Fungus Example

MIM (Edwards, 2000) selected optimal model using a greedy stepwise backward model selection procedure based on testing individual edges.

The final model is biologically plausible.

Parameter Estimability: Mildew Fungus Example

Sequence of models found by `MIM`. Red boxes indicate zero margins (MLE does not exist).

Model	Unadjusted d.f.	Adjusted d.f.
[ABCDEF]	0	0
[ABCEF] [ABCDE]	16	3
[BCEF] [ABCDE]	24	6
[BCEF] [ABCE] [ABCD]	32	12
[BCEF] [ABCE] [ABD]	36	17
[BCEF] [AD] [ABCE]	38	18
[CEF] [AD] [ABCE]	42	22
[CEF] [AD] [BCE] [ABE]	46	27
[CEF] [AD] [ABE]	48	29
[CEF] [AD] [BE] [AB]	50	31
[CF][CE][AD] [BE] [AB]	52	37

Extended Maximum Likelihood Estimation: Numerical Procedures

How to compute the extended MLE?

Extended Maximum Likelihood Estimation: Numerical Procedures

How to compute the extended MLE?

If $\mathrm{A}^\top n \in \mathrm{ri}(F)$, for some face of C_A, then

- (non-zero) points in the normal cone to F are directions of recession of the negative log-likelihood;
- the Fisher information $I(\theta)$ matrices for $\mathcal{E}_F$ are singular.

$$\zeta^\top I(\theta)\zeta = 0, \quad \forall \zeta \in \mathcal{L}_F, \forall \theta \in \mathbb{R}^d.$$

Extended Maximum Likelihood Estimation: Numerical Procedures

How to compute the extended MLE?

If $\mathrm{A}^\top n \in \mathrm{ri}(F)$, for some face of C_A, then

- (non-zero) points in the normal cone to F are directions of recession of the negative log-likelihood;
- the Fisher information $I(\theta)$ matrices for $\mathcal{E}_F$ are singular.

$$\zeta^\top I(\theta)\zeta = 0, \quad \forall \zeta \in \mathcal{L}_F, \forall \theta \in \mathbb{R}^d.$$

Two-step procedure:

- Step 1
  Identify the facial set $\mathcal{F}$ and compute a basis for $\pi_{\mathcal{F}}\left(\mathcal{M} \cap (\mathcal{N} + \mathcal{L}_F)^c\right)$.
  To compute $\mathcal{F}$:

  Given $\mathbf{t} = \mathrm{A}^\top n$*, determine the facial set* $\mathcal{F}$ *of rows of* A *which span the face* F *of* C_A *such that* $\mathbf{t} \in \mathrm{ri}(F)$*.*

Extended Maximum Likelihood Estimation: Numerical Procedures

How to compute the extended MLE?

If $\mathrm{A}^\top n \in \mathrm{ri}(F)$, for some face of $\mathrm{C_A}$, then

- (non-zero) points in the normal cone to F are directions of recession of the negative log-likelihood;
- the Fisher information $I(\theta)$ matrices for $\mathcal{E}_F$ are singular.

$$\zeta^\top I(\theta)\zeta = 0, \quad \forall \zeta \in \mathcal{L}_F, \forall \theta \in \mathbb{R}^d.$$

Two-step procedure:

- Step 1
  Identify the facial set $\mathcal{F}$ and compute a basis for $\pi_{\mathcal{F}}\left(\mathcal{M} \cap (\mathcal{N} + \mathcal{L}_F)^c\right)$.
  To compute $\mathcal{F}$:

  Given $\mathbf{t} = \mathrm{A}^\top n$, determine the facial set $\mathcal{F}$ of rows of A which span the face F of C_{A} such that $\mathbf{t} \in \mathrm{ri}(F)$.

- Step 2
  Optimize the restricted likelihood function for the extended family.
  When $\mathcal{F}$ is available, this is equivalent to maximizing the likelihood of a log-linear model with structural zeros along $\mathcal{F}$. (Easy)

Algorithms for Extended Maximum Likelihood Estimation

Step 1 (finding $\mathcal{F}$) is the important one.

- Let A_+ and A_0 the sub-matrix of A corresponding to positive and zero entries of *n*.

Algorithms for Extended Maximum Likelihood Estimation

Step 1 (finding $\mathcal{F}$) is the important one.

- Let A_+ and A_0 the sub-matrix of A corresponding to positive and zero entries of *n*.
  Identifying $\mathcal{F}^c$ is equivalent to finding of a vector $c \in \mathbb{R}^d$ such that:
  - $\mathrm{A}_+\mathbf{c} = \mathbf{0}$;
  - $\mathrm{A}_0\mathbf{c} \geqslant \mathbf{0}$;
  - the set $\mathrm{supp}(\mathrm{A}\mathbf{c})$ has maximal cardinality.

Algorithms for Extended Maximum Likelihood Estimation

Step 1 (finding $\mathcal{F}$) is the important one.

- Let A_+ and A_0 the sub-matrix of A corresponding to positive and zero entries of *n*.
  Identifying $\mathcal{F}^c$ is equivalent to finding of a vector $c \in \mathbb{R}^d$ such that:
  - $\mathrm{A}_+\mathbf{c} = \mathbf{0}$;
  - $\mathrm{A}_0\mathbf{c} \geqslant \mathbf{0}$;
  - the set $\mathrm{supp}(\mathrm{A}\mathbf{c})$ has maximal cardinality.
- This can be done with repeated iterations of LP (see also Geyer, 2009) or with non-linear methods.
  For large problems, they can be computationally intensive.

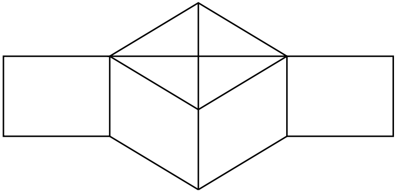
Relevance to exact inference: knowledge of $\mathcal{F}$ can help finding Markov bases.

Reducible Hierarchical Log-Linear Models

- For reducible hierarchical log-linear models both tasks can be carried out in parallel over appropriate sub-models.
- A hierarchical log-linear model (simplicial complex) is reducible if it can be obtained as the direct join of two sub simplicial complex (see Lauritzen, 1996). Apply the definition recursively.

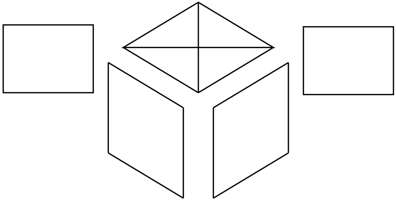
Reducible Hierarchical Log-Linear Models

- For reducible hierarchical log-linear models both tasks can be carried out in parallel over appropriate sub-models.
- A hierarchical log-linear model (simplicial complex) is reducible if it can be obtained as the direct join of two sub simplicial complex (see Lauritzen, 1996). Apply the definition recursively.



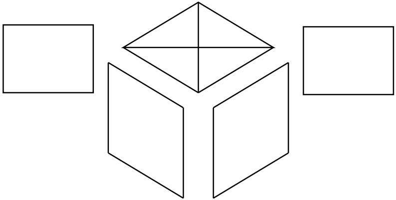
Reducible Hierarchical Log-Linear Models

- For reducible hierarchical log-linear models both tasks can be carried out in parallel over appropriate sub-models.
- A hierarchical log-linear model (simplicial complex) is reducible if it can be obtained as the direct join of two sub simplicial complex (see Lauritzen, 1996). Apply the definition recursively.



Reducible Hierarchical Log-Linear Models

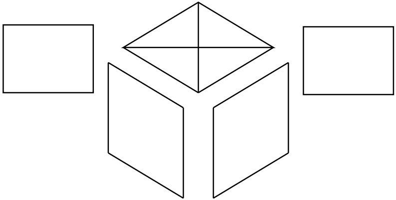
- For reducible hierarchical log-linear models both tasks can be carried out in parallel over appropriate sub-models.
- A hierarchical log-linear model (simplicial complex) is reducible if it can be obtained as the direct join of two sub simplicial complex (see Lauritzen, 1996). Apply the definition recursively.



- Theoretical justification: reducible models are defined by cuts (Barndorff-Nielsen, 1974).

Reducible Hierarchical Log-Linear Models

- For reducible hierarchical log-linear models both tasks can be carried out in parallel over appropriate sub-models.
- A hierarchical log-linear model (simplicial complex) is reducible if it can be obtained as the direct join of two sub simplicial complex (see Lauritzen, 1996). Apply the definition recursively.



- Theoretical justification: reducible models are defined by cuts (Barndorff-Nielsen, 1974).
- Old idea: Hara et al. (2011), Engsrtöm et al. (2011), Sullivant (2007), Eriksson et al. (2006), Dobra and Sullivant (2004), Badsberg and Malvestuto (2001), Frydenberg (1990), Leimer (1993), Tarjan (1985).

Still lots of work ahead...

- The validity/applicability of model selection based on adjusted degree of freedom has to be investigated, especially in the double asymptotic framework.

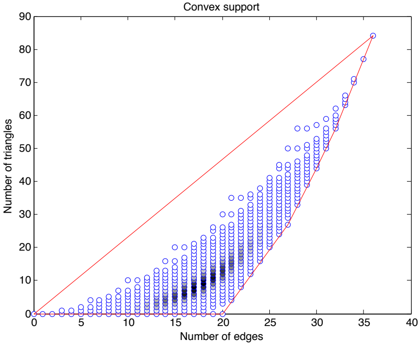
- Computationally efficient methods for model selection for large tables still lacking.

Still lots of work ahead...

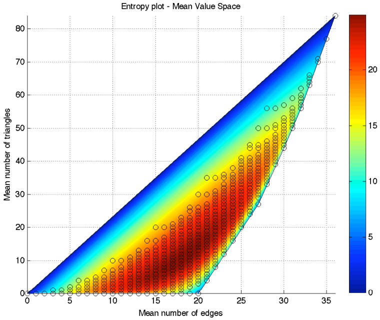
Thank you

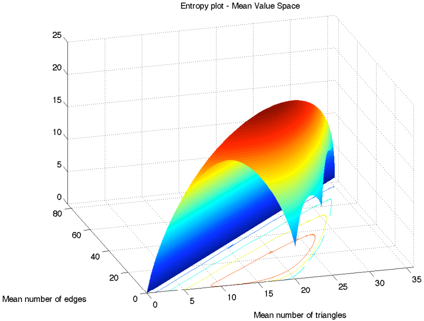
- Fienberg S. E. and Rinaldo, A. (2011). Maximum Likelihood Estimation in Log-linear Models, http://arxiv.org/abs/1104.3618
- Rinaldo, A., Petrović, S. and Fienberg, S. E. (2011). Maximum Likelihood Estimation in Network Models, http://arxiv.org/abs/1105.6145
- Rinaldo, A., Fienberg, S. E. and Zhou, Y. (2009). On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models, *Electronic Journal of Statistics,* 3, 446–484.

9 nodes: sufficient statistics are the number of edges and triangles.

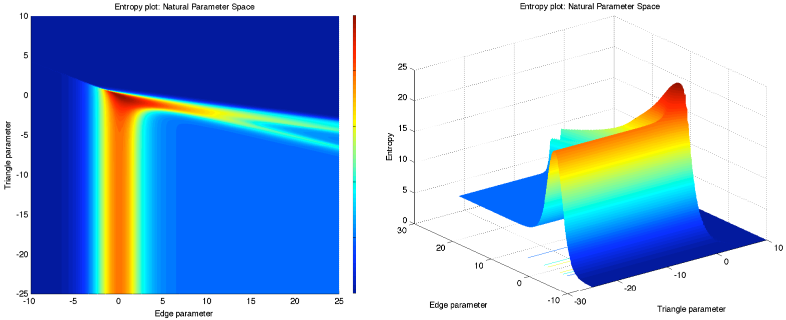


Entropy plot over the mean value space.

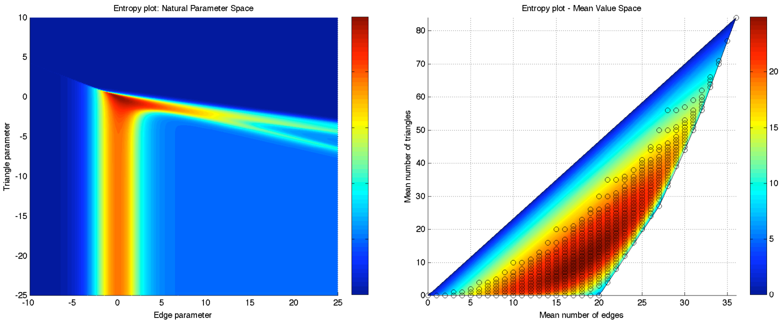




Entropy plot over the natural parameter space.



Entropy plots of the natural space and mean value spaces.



Entropy plots of the natural space space with superimposed the normal fan and of the mean value space.