Hierarchical statistical models and special factorizations

Henry Wynn

Contents

- Conditional independence and other factorizations
- Polynomial models
- Monomial ideals and Betti numbers
- Experimental design
- Aberration and corner cut models
- Lattice conditional independence models
- A new class of models based on networks
- A way forward

Conditional independence and other factorizations

Let $X_1, \ldots, X_n$ be set of jointly distributed univariate random variables.

Let $\mathbb{N} = \{1, \ldots, n\}$ and for any subsets of index of index set $I \subseteq \mathbb{N}$, we let $f_J(x_J)$ be the marginal joint density function of $X_J = \{x_j,\ i \in J\}$, which we shall assume are always positive and continuous. For non empty index set $I, J \subset \mathbb{N}$ we define conditional independence of X_I and X_J give $X_{I \bigcap J}$ by

$$f_{I \cup J}(x_{I \cup J}) = \frac{f_I(x_I) f_J(x_J)}{f(x_{I \bigcap J})},$$

for all x.

Another factorization

Assume that $I \cap J \cap K = \emptyset$ then take

$$f_{I\cup J\cup K}(x_{I\cup J\cup K}) = \frac{f_{I\cup J}(x_{I\cup J}) f_{I\cup K}(x_{I\cup K}) f_{J\cup K}(x_{J\cup K})}{f_{I\cap J}(x_{I\cap J}) f_{I\cap K}(x_{I\cap K}) f_{J\cap K}(x_{J\cap K})}$$

For $I = \{1,2\},\ J = \{1,3\},\ K = \{2,3\}$ this is

$$f_{123} = \frac{f_{12} f_{13} f_{23}}{f_1 f_2 f_3}$$

or in the discrete case

$$p_{ijk} = \frac{p_{ij\cdot}\ p_{i\cdot k}\ p_{\cdot jk}}{p_{i\cdot\cdot}\ p_{\cdot j\cdot}\ p_{\cdot\cdot k}}$$

In the binary case this is equivalent to toric ideal

$$p_{000} p_{110} p_{101} p_{001} - p_{100} p_{010} p_{001} = 0$$

Polynomial models

Regression:

$$\eta(x_1, x_2, x_3) = \theta_{000} + \theta_{100}x_1 + \theta_{010}x_2 + \theta_{001}x_3$$
$$+\theta_{110}x_1x_2 + \theta_{101}x_1x_3$$

Conditional independence for log-linear models:

$$\log p(x) = \theta_{000} + \theta_{100}x_1 + \theta_{010}x_2 + \theta_{001}x_3 + \theta_{101}x_1x_2 + \theta_{011}x_1x_3$$

Other example

$$\log p(x) = \theta_{000} + \theta_{100}x_1 + \theta_{010}x_2 + \theta_{001}x_3$$
$$+\theta_{110}x_1x_2 + \theta_{101}x_1x_3 + \theta_{011}x_2x_3$$

Both these are *hierarchical* models. Equivalently they are based on *simplicial complexes*. Also: *staircase* models.

Models and toric ideals

Models are independent of the design/support. Toric varieties/ideals are dependent on model *and* the support (usually a product space).

Many examples of same model but different designs: incomplete layouts, orthogonal fractions, balanced incomplete designs, response surface designs, optimal designs....

Monomial ideals

A model simplicial complex has a one-to-one relation with a special type of monomial ideal, called the Stanley-Reisner ideal.

For a simplicial complex Δ, let I_Δ be the squarefree monomial ideal created by the non-faces of Δ: $I_\Delta = \langle x^a : a \notin \Delta \rangle$.

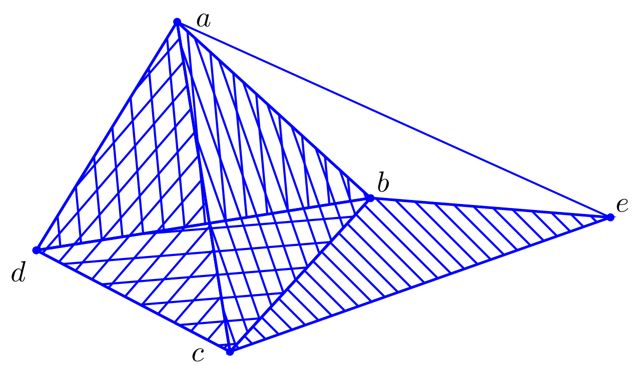
The complexity of the model Δ can be studied by the Stanley-Reisner ring $R[x]/I_\Delta$.

In the description of $R[x]/I_\Delta$, Betti numbers play a central role. Graded Betti number is the minimal number of generators $e_{a,j}$ in degree $i + j$.

Betti numbers

Consider the following simplicial complex (model): $\Delta = \{1, a, b, c, d, e, ab, ac, ad, ae, bc, bd, be, cd, ce, abc, abd, adc, bcd, bce\}$ and the Stanley-Reisner ideal generated by it:

$$I_\Delta = \langle de, abe, ace, abcd \rangle \subset \mathbb{R}[a, b, c, d, e]$$



```
Use T::=Q[a,b,c,d,e]; J:=Ideal(de,abe,ace,abcd); Hill
HilbertSeries(T/J);

H(t) = 5/2t^2 + 3/2t + 1   for t >= 0
-------------------------------
(1 + 2a + 2a^2) / (1-a)^3
-------------------------------
```

```
BettiDiagram(T/J);

        0     1     2     3
--------------------------
 0:     1     -     -     -
 1:     -     1     -     -
 2:     -     2     3     1
 3:     -     1     1     -
--------------------------
```

Hilbert Series (after simplification)

$$HS = \frac{\begin{pmatrix} 1 - de - abe - ace - abcd \\ +abde + acde + abce \end{pmatrix}}{(1-a)(1-b)(1-c)(1-d)(1-e)}$$

```
BettiDiagram(J);

        0    1    2
--------------------
 2:     1    -    -
 3:     2    3    1
 4:     1    1    -
--------------------
Tot:    4    4    1
---------------------
```

Alternatively we could use the Artinian closure of $\bar{I}_\triangle$ which is

$$I_\triangle = \langle de, abe, ace, abcd\rangle + \langle a^2, b^2, c^2, d^2, e^2\rangle$$

```
K:=J+Ideal(a^2,b^2,c^2,d^2,e^2); Hilbert(T/K); Hilber
-------------------------------
H(0) = 1 H(1) = 5 H(2) = 9 H(3) = 5 H(t) = 0   for t
-------------------------------
(1 + 5a + 9a^2 + 5a^3)
-------------------------------
```

```
K; BettiDiagram(T/K);

Ideal(de, abe, ace, abcd, a^2, b^2, c^2, d^2, e^2)
-------------------------------
         0    1    2    3    4
------------------------------
 2:      6    2    -    -    -
 3:      2   21   21    7    1
 4:      1    5   17   17    5
------------------------------
Tot:     9   28   38   24    6
-------------------------------
```

A degree-by-degree description of the model border is obtained.

A example of a design

Consider a Plackett-Burman (PB8) design with eight runs, seven factors a, b, c, d, e, f, g and generator $+--+-++$.

a	b	c	d	e	f	g
1	-1	-1	1	-1	1	1
1	1	-1	-1	1	-1	1
1	1	1	-1	-1	1	-1
-1	1	1	1	-1	-1	1
1	-1	1	1	1	-1	-1
-1	1	-1	1	1	1	-1
-1	-1	1	-1	1	1	1
-1	-1	-1	-1	-1	-1	-1

The equivalence classes of models

The results for the 218 models in the algebraic fan of PB8 are summarized in Table 2, where representatives of six equivalence classes (up to permutation of factors) are shown.

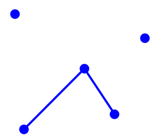
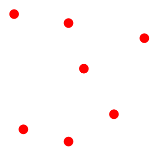
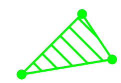
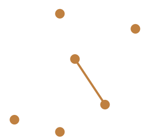
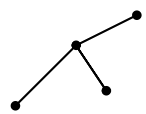
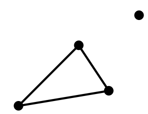
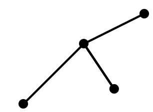
 $1+5s+2s^2$ (84)	$\prec_2$ (DegRevLex)  $1+7s$ (1)	$\prec_3$ (Lex)  $1+3s+3s^2+s^3$ (28)
 $1+6s+s^2$ (21)	$\prec_1$ (Block)  (56) $1+4s+3s^2$ (28)	

Table 2: Equivalence classes of models Δ and corresponding Hilbert Series for PB8.

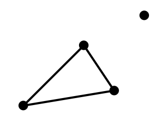
PB8: Comparing models

Betti table									Model

	0	1	2	3	4	5	6	7
0:	1	3	3	1	-	-	-	-
1:	-	7	29	48	40	17	3	-
2:	-	-	6	26	45	39	17	3
Tot:	1	10	38	75	85	56	20	3

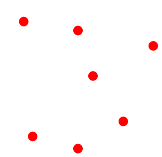


	0	1	2	3	4	5	6	7
0:	1	3	3	1	-	-	-	-
1:	-	7	30	52	47	24	7	1
2:	-	1	10	33	52	43	18	3
Tot:	1	11	43	86	99	67	25	4



Disconnected model $\prec_2$ `DegRevLex`:

	0	1	2	3	4	5	6
0:	1	-	-	-	-	-	-
1:	-	21	70	105	84	35	6
Tot:	1	21	70	105	84	35	6



Matching statistical models with monomial ideals

To repeat: model simplicial complex (hierarchical model) has a one-to-one relation with a special the Stanley-Reisner ideal. This relationship is the same whether we consider *regression* or *log-linear binary categorical model*. But the interpretation is different.

- Decomposable graphical models
- Corner-cut models
- Lattice conditional Independence (LCI) models
- New classes of models

Decomposabled graphical models

Known results.

- Definition of decomposability
- Chordal
- Is a junction tree
- Quadratic toric ideal

Note that the first three are model statements. But what are the results for monomial ideals?

- $I_\triangle$ has a "2-resolution"
- Recent results on Betti numbers

Linear aberration

- Taking the motivation from the concept of aberration, we want to fill out lower degrees before higher:

$$A(w, L) = \frac{1}{n} \sum w_i \bar{\alpha}_{L_i}$$

$w_i \geq 0$, $\sum w_i = 1$.

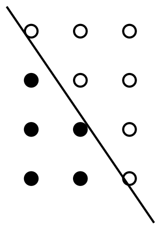
Theorem: Any algebraic model minimises some $A(w, L)$.

Proof. Use LP arguments for the lower boundary of $\mathcal{S}(I)$.

- **Generic** designs minimise $A(w, L)$ over $\mathcal{C}_{d,n}$ and all vectors w.
- For generic designs, algebraic models are corner cut models.

Corner cut models

Corner-cut model



$\{1, x_1, x_2, x_1 x_2\}$ is not corner cut

Corner-cut models and linear aberration

- The state polytope summarises information about linear aberration, i.e. its vertexes correspond to models that minimise $A(w, L)$ over the set of identifiable hierarchical models S.

- The vertexes of $\mathcal{S}(I)$ correspond to algebraic models A.

- The (minimum) aberration of designs can be compared through their state polytopes.

- However, there may be non-algebraic models on the lower boundary (and thus minimising $A(w, L)$ for some w) or in the interior of $\mathcal{S}(I)$.

Maximum Betti numbers and corner cut

Bigatti-Hulett-Pardue Theorem: Let $I \subset R$ an ideal and let L be the lex ideal such that $\mathcal{H}(R/I) = \mathcal{H}(R/L)$. Then $\beta^L_{i,j} \geq \beta^I_{i,j}$ for all $i = 1, \ldots, n$ and $j \in \mathbb{N}$.

"Lex segment" construction

New result: For fixed Hilbert function the ideals satisfying BHP theorem are generalized corner cut (intersection of the state polytope with the hyperplane of fixed Hilbert function) and have a vertex in the state polytope which is on the lower convex boundary of all models with that Hilbert function.

Note for 2^k factorial independence models are CC, but (conjecture CI and other models may be generalised CC.

Models and factorisations

Consider, for each x the log-density as a function on the index set:

$$\lambda(J) = \log f_J(x_J),$$

so that conditional independence is equivalent to

$$\lambda(I \cup J) = \lambda(J) + \lambda(J) - \lambda(I \cap J) \tag{1}$$

If we add the condition that $\lambda(\emptyset) = 0$ we can also include independence: when $I \cap J = 0$:

$$\lambda(I \cup J) = \lambda(J) + \lambda(J)$$

Condition (1) is the condition for a *valuation*.

Generally additive valuations

Begin with a collection of index sets set $\mathcal{S}_0$, closed under intersection. A valuation μ on $\mathcal{S}_0$ is generally additive if for all any $I_1, \ldots, I_m \in \mathcal{S}_0$ with $I_1 \cup I_2 \cdots \cup I_m \in \mathcal{S}$

$$\mu(I_1 \cup I_2 \cdots \cup I_m) = \sum_i \mu(I_i) - \sum_{i<j} \mu(I_i \cap I_j)$$

$$+ \sum_{i<j<k} \mu(I_i \cap I_j \cap I_k) \cdots + (-1)^{m-1} \mu(\cap_{i=1}^m I_i)$$

Groemer-Perles-Sallee extension theorem: the following are equivalent

(i) μ is generally additive.

(ii) μ has an additive extension to finite unions.

Simplified factorisations

(i) Assume general additivity of λ: Lattice Conditional independence, also equivalent to TDAG.

(ii) Seek complexity reductions in inclusion-exclusion relations using generalisation of junction tree.

Special case: junction tree

Example. Let $n = 5$

$$\mathcal{S}_0 = \{I_1, I_2, I_3\} = \{\{1,2,3\},\{2,3,4\},\{3,4,5\}\}$$

The full *depth* 3 inclusion-exclusion identity is

$$\begin{aligned} \lambda(\{1,2,3,4,5\}) &= \lambda(\{1,2,3\} \cup \{2,3,4\} \cup \{3,4,5\}) \\ &= \lambda(\{1,2,3\}) + \lambda(\{2,3,4\}) + \lambda(\{3,4,5\}) \\ &\quad -\lambda(\{2,3\}) - \lambda(\{3,4\}) \\ &\quad -\lambda(\{3\}) + \lambda(\{3\}). \end{aligned}$$

$$\lambda(\{1,2,3,4,5\}) = \lambda(\{1,2,3\}) + \lambda(\{2,3,4\}) + \lambda(\{3,4,5\}) - \lambda(\{2,3\}) - \lambda(\{3,4\}).$$

The set-theoretic condition which gives the cancelation is running intersection property that

$$(I_1 \cap I_3) \subset I_2$$

TDAG representation

$$\begin{array}{ccc} 2 & \rightarrow & 1 \\ \uparrow & \nearrow & \\ 3 & & \\ \downarrow & \searrow & \\ 4 & \rightarrow & 5 \end{array}$$

Generating sets correspond to maximal chains.

Conditional expectations as valuations

$$\mathbb{E}(\cdot|X_J)$$

$$\mathbb{E}(\cdot|X_{I\cup J\cup K}) = \mathbb{E}(\cdot|X_I) + \mathbb{E}(\cdot|X_J) - \mathbb{E}(\cdot|X_{I\cap J})$$

Theorem . Let λ be generally additive on a intersection closed class of sets $\mathcal{S}_0$. Let

$$\mathcal{H}(\mathcal{S}_0) = \{h(x) = \sum_{I\in\mathcal{S}_0} g_I(x_I)\}$$

$$\mu_h(I) = \mathbb{E}(h(x)|x_I),$$

is a generally additive valuation.

One way of thinking about this is that we have a natural set of regressions.

Commutivity: Gaussian versus the rest

The Gaussian case is particularly simple. Think of MA representation:

$$X_I = A_I \varepsilon$$

Then to every X_I there is a subspace and a projector P_I onto that subspace. Pairwise additivity implies general additivity

$$P_{I \cup J} = P_I + P_J - P_{I \cap J} = P_I + P_J - P_I P_J$$

Equivalent to commutativity!

$$P_I P_J = P_J P_I$$

Equivalent to the same for conditional expectations:

$$\mathbb{E}_I \mathbb{E}_J = \mathbb{E}_J \mathbb{E}_I$$

For the non-Gaussian case the properties hold *on* $H(\mathcal{S}_0)$.

Because of commutativity we have an understanding of LCI in the Gaussian case:

(i) All the covariance matrices C_I in the full Boolean lattice are simulanteously diagonalizable

(ii) In the diagonal form the projectors are diagonal with entries $0, 1$. eg

$$I_{3\times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; P_{12} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$P_{13} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$I = P_{12} + P_{13} - P_1, \; P_{12}P_{13} = P_{13}P_{12}$$

Boolean operations: $(1,1,1) = (1,1,0) + (1,0,1) - (1,0,0)$

Discrete case

Change in notation:

$$\lambda(I) = \log p_I$$

where p_I is the marginal probability. eg CI is

$$\log p_{ijk} = \log p_{ij\cdot} + \log p_{i\cdot k} - \log p_{i\cdot\cdot}$$

Conditional expectations:

$$\mathbb{E}_{I\cup J} = \mathbb{E}_J + \mathbb{E}_J - \mathbb{E}_{I\cap J}$$

on a restricted subspace (details omitted).

Shannon information

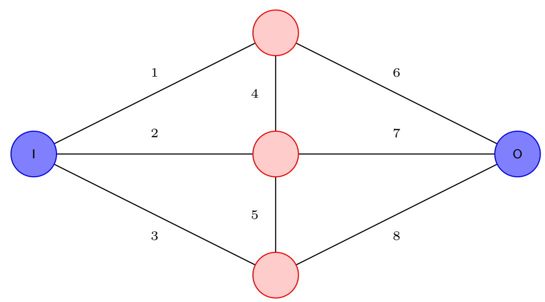
For any junction tube we have all corresponding identities for Shannon information:

$$\mathcal{I}(I) = \mathbb{E}(\lambda(I))$$

$$\mathcal{I}(I_1 \cup I_2 \cdots \cup I_m) = \sum_1 \mathcal{I}(I_i) - \sum_2 \mathcal{I}(I_i \cap I_j) + \cdots + (-1)^d \sum_d \mathcal{I}(\bigcap I_i)$$

New classes of models from “new” monomial ideals

One class: networks



Paths give generators of I_Δ:

$$I_\Delta =< x_1x_6, x_1x_4x_7, x_2x_4x_6, x_1x_4x_5x_8, x_2x_7, \\ x_3x_4x_5x_6, x_2x_5x_8, x_3x_5x_7, x_3x_8\rangle$$

Model cliques:

$$234578, 134567, 124567, 23568, 13578, 13467, 12468, 2367, 1278$$

The future

- Model complex versus SR monomial ideal: I_Δ
- Betti numbers of I_Δ and minimal free resolution (and the model complex)
- Use Hochster's formula to link to Betti number of the model to those of I_Δ
- Factorizations of the full joint pdf
- New models by starting with I_Δ, or Artinian closure

One way forward: junction tubes?

Under certain conditions $\mathcal{C}$ on the set of cliques the minimal free resolution of $I_\triangle$ and will yield the "optimal" factorization(s) of the joint probability.

1. Start with square free monomial ideal $(I_\triangle)$
2. From this we get a set of model cliques
3. Assume that corresponding log-likelihood is generally additive on the Boolean lattice formed by the cliques: (LCI model)
4. This corresponds to a factorization based on the Taylor resolution (simple inclusion-exclusion)

5. Under special conditions $\mathcal{C}$ we can reduce the complexity of the factorization

6. One choice for $\mathcal{C}$ is a generalization of the running intersection property

7. (Hope) This is the first large and interesting class of monomial ideals and includes decomposable graphic models as a special case

8. Find some interpretation of the Betti numbers in terms of factorizations of probability

9. Link the projection representation to the algebra

Papers

1. Sonya Petrovich and Erik Stokes: Markov degrees..... (very useful)

2. Bernstein, Y., Maruri-Aguilar, H., Shmuel O, Riccomagno, E and HPW: Annals of the Institute of Statistical Mathematics, 62 (2010),

3. Hugo Maruri-Aguilar, Eduardo Sáenz de Cabezon and HPW: paper on Betti numbers in experimental design (nearly finished)

4. Gianni Pistone, Jim Smith, Eduardo S-de-C and HPW: paper on Junction Tubes (been writing for a few years!)

5. Daniel Bruynooghe (LSE PhD Student) and HPW: new paper on hierarchical models.