The model space

Matthew Macauley

Department of Mathematical Sciences
Clemson University
http://www.math.clemson.edu/~macaule/

Algebraic Biology

Inferring Boolean models from partial data

Suppose a Boolean model (f_1, f_2, f_3) has the following (partial) state space (s_i, s_i).

$$001 \longrightarrow 101 \longrightarrow 111 \longrightarrow 110 \longrightarrow 010 \longrightarrow 000 \qquad 100 \qquad 011$$

Main question

What are the possible Boolean models (f_1, f_2, f_3), as polynomials?

This is **stronger** than what we asked last time: *which variables depend on which variables?*

x	0	0	1	1	0	0	1	1
y	0	1	0	1	0	1	0	1
z	0	0	0	0	1	1	1	1
$f_1(x,y,z)$	?	0	?	0	1	?	1	1
$f_2(x,y,z)$	?	0	?	1	0	?	1	1
$f_3(x,y,z)$	?	0	?	0	1	?	1	0

As before, we can treat each function f_1, f_2, f_3 separately.

First question

What are the possible functions $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$, given partial information?

A familiar example

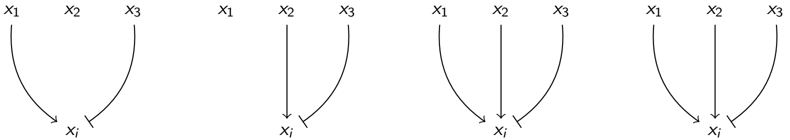
Recall the following unknown Boolean function:

$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f_i(x)$	0	1	?	?	?	?	?	0

Of the 256 Boolean functions on 3 variables, $2^{8-3} = 32$ fit this data, and only 4 are unate:

$$x_1 \wedge \overline{x_3}, \qquad x_2 \wedge \overline{x_3}, \qquad x_1 \wedge x_2 \wedge \overline{x_3}, \qquad (x_1 \vee x_2) \wedge \overline{x_3}.$$

The wiring diagrams of these functions are shown below, expressed several different ways.



$(1, 0, -1)$ $\qquad$ $(0, 1, -1)$ $\qquad$ $(1, 1, -1)$ $\qquad$ $(1, 1, -1)$

$\{x_1, \overline{x_3}\}$ $\qquad$ $\{x_2, \overline{x_3}\}$ $\qquad$ $\{x_1, x_2, \overline{x_3}\}$ $\qquad$ $\{x_1, x_2, \overline{x_3}\}$

This time, we'll find the actual functions, in polynomial form.

A familiar example

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

First step: interpolation

Find a single function $f\colon \mathbb{F}_2^3 \to \mathbb{F}_2$ that fits the data.

For each data point s_i, we'll construct an *r*-polynomial that has the following property:

$$r_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} = \mathrm{s}_j\, j \neq i \\ 0 & \mathrm{x} = \mathrm{s}_j,\ j \neq i \end{cases}$$

Once we have these, one such polynomial $f(\mathrm{x})$ we seek will be

$$f(\mathrm{x}) = t_1 r_1(\mathrm{x}) + t_2 r_2(\mathrm{x}) + t_3 r_3(\mathrm{x}).$$

Note *why* this works:

$$f(\mathrm{s}_1) = t_1 r_1(\mathrm{s}_1) + t_2 r_2(\mathrm{s}_1) + t_3 r_3(\mathrm{s}_1) = t_1 \cdot 1 + t_2 \cdot 0 + t_3 \cdot 0 = t_1$$
$$f(\mathrm{s}_2) = t_1 r_1(\mathrm{s}_2) + t_2 r_2(\mathrm{s}_2) + t_3 r_3(\mathrm{s}_2) = t_1 \cdot 0 + t_2 \cdot 1 + t_3 \cdot 0 = t_2$$
$$f(\mathrm{s}_3) = t_1 r_1(\mathrm{s}_3) + t_2 r_2(\mathrm{s}_3) + t_3 r_3(\mathrm{s}_3) = t_1 \cdot 0 + t_2 \cdot 0 + t_3 \cdot 1 = t_3$$

A familiar example: $k = 1$

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

For each data point s_i, we'll construct an r-**polynomial**, satisfying

$$r_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} = \mathrm{s}_i \\ 0 & \mathrm{x} = \mathrm{s}_j,\ j \neq i \end{cases}$$

One function that works is

$$r_i(\mathrm{x}) = \prod_{\substack{k=1 \\ k \neq i}}^{m} (x_{\ell_k} - s_{k\ell_k})$$

where ℓ_k is *any* coordinate in which s_i and s_k differ. To construct $r_1(x)$ from this example:

- $k = 2$: use $x_3 - 0$
- $k = 3$: use $x_1 - 0$, $x_2 - 0$, or $x_3 - 0$.

Thus, we can use any of the following for $r_1(x)$:

$$r_1(x) = x_1x_3, \qquad r_1(x) = x_2x_3, \qquad \text{or} \qquad r_1(x) = x_3^2 = x_3,$$

(Since we only care about functions, we may reduce $x_i^2 = x_i$.)

A familiar example $k = 2$

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

For each data point s_i, we'll construct an r-**polynomial** that satisfies

$$r_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} = \mathrm{s}_i \\ 0 & \mathrm{x} = \mathrm{s}_j \, j \neq i \end{cases}$$

One function that works is

$$r_i(\mathrm{x}) = \prod_{\substack{k=1 \\ k \neq i}}^{m} (x_{\ell_k} - s_{k\ell_k})$$

where ℓ_k is *any* coordinate in which s_i and s_k differ. To construct $r_2(x)$ from this example,

- $k = 1$: use $x_3 - 1$
- $k = 3$: use $x_1 - 0$ or $x_2 - 0$.

Thus, we can use any of the following for $r_2(x)$:

$$r_2(x) = x_1(x_3 + 1), \qquad \text{or} \qquad r_2(x) = x_2(x_3 + 1).$$

A familiar example $k = 3$

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

For each data point s_i, we'll construct an r-**polynomial** that satisfies

$$r_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} = \mathrm{s}_i \\ 0 & \mathrm{x} = \mathrm{s}_j,\ j \neq i \end{cases}$$

One function that works is

$$r_i(\mathrm{x}) = \prod_{\substack{k=1 \\ k \neq i}}^{m} (x_{\ell_k} - s_{k\ell_k})$$

where ℓ_k is *any* coordinate in which s_i and s_k differ. To construct $r_3(x)$ from this example,

- $k = 1$: use $x_1 - 1$, $x_2 - 1$, or $x_3 - 0$.
- $k = 2$: use $x_1 - 1$ or $x_2 - 1$.

Thus, we can use any of the following for $r_3(x)$:

$$r_3(x) = (x_1 + 1), \qquad r_3(x) = (x_2 + 1), \qquad r_3(x) = (x_1 + 1)(x_2 + 1),$$

$$r_3(x) = (x_1 + 1)x_3 \qquad \text{or} \qquad r_3(x) = (x_2 + 1)x_3.$$

The vanishing ideal

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

One such choice for r_1, r_2, and r_3 yields

$$\begin{aligned} f(\mathrm{x}) &= t_1 r_1(\mathrm{x}) + t_2 r_2(\mathrm{x}) + t_3 r_3(\mathrm{x}) \\ &= 0 \cdot x_3 + 1 \cdot x_1(x_3+1) + 0 \cdot (x_2+1) \\ &= x_1(x_3+1). \end{aligned}$$

We just found a single function that fits the data. Now, let's find *every* such function.

Proposition

Let $f(\mathrm{x}) \in \mathbb{F}[x_1,\dots,x_n]/\langle x_1^2-x_1,\dots,x_n^2-x_n\rangle$ fit a set $\mathcal{D} = \{(\mathrm{s}_1,t_1),\dots,(\mathrm{s}_k,t_k)\}$ of **data**.

(i) If $h(\mathrm{x})$ **vanishes** on all s_i, then $f(\mathrm{x})+h(\mathrm{x})$ fits the data.

(ii) The polynomials that vanish on the data form an ideal $I(\mathcal{D})$.

(iii) Every polynomial $g(x)$ that fits the data can be written as $g(\mathrm{x}) = f(\mathrm{x}) + h(\mathrm{x})$ for some $h(\mathrm{x}) \in I(\mathcal{D})$.

The structure of the model space

Theorem / Definition

Consider a set $\mathcal{D} = \big\{(\mathrm{s}_1, t_1), \dots, (\mathrm{s}_k, t_k)\big\}$ of **data**, where $\mathrm{s}_i \in \mathbb{F}^n$, $\mathrm{t}_i \in \mathbb{F}$, and $|\mathbb{F}| = q$.

The set of functions that fit the data is the model space

$$\mathrm{Mod}(\mathcal{D}) := f + I(\mathcal{D}) = \big\{f + h \mid h \in I(\mathcal{D})\big\},$$

where f is *any* function that fits the data, and $I(\mathcal{D})$ is the vanishing ideal in

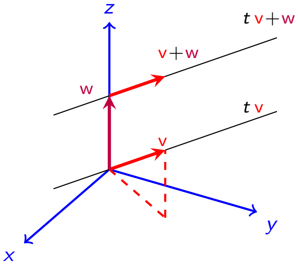
$$\mathbb{F}[x_1, \dots, x_n]/\langle x_1^q - x_1, \dots, x_n^q - x_n\rangle,$$

Here are some other mathemtical problems whose solutions have a similar structure.

1. Parametrize a line in $\mathbb{R}^n$.
2. Parametrize a plane in $\mathbb{R}^n$.
3. Solve the underdetermined system $\mathrm{Ax} = \mathrm{b}$.
4. Solve the differential equation $x'' + x = 2$.

Parametrize a line in $\mathbb{R}^n$

Suppose we want to write the equation for a line that contains a vector $\mathrm{v} \in \mathbb{R}^n$:



This line, which *contains the zero vector*, is $t\mathrm{v} = \{t\mathrm{v} : t \in \mathbb{R}\}$.

Now, what if we want to write the equation for a line parallel to v?

This line, which *does not contain the zero vector*, is

$$t\mathrm{v} + \mathrm{w} = \{t\mathrm{v} + \mathrm{w} : t \in \mathbb{R}\}.$$

Note that ANY particular w on the line will work!!!

Solve an underdetermined system $Ax = b$

Suppose we have a system of equations that has "too many variables," so there are infinitely many solutions.

For example:

$$\begin{aligned} 2x + y + 3z &= 4 \\ 3x - 5y - 2z &= 6 \end{aligned} \qquad \text{“}Ax = b\text{ form”:} \qquad \begin{bmatrix} 2 & 1 & 3 \\ 3 & -5 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}.$$

How to solve:

1. Solve the related homogeneous equation $Ax = 0$ (this is null space, $\mathrm{NS}(A)$);
2. Find any particular solution x_p to $Ax = b$;
3. Add these together to get the general solution: $x = \mathrm{NS}(A) + x_p$.

This works because geometrically, the solution space is just a line, plane, etc.

Here are two possible ways to write the solution:

$$C \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \qquad C \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 10 \\ 8 \\ -8 \end{bmatrix}.$$

Linear differential equations

Solve the differential equation $x'' + x = 2$.

How to solve:

1. Solve the related homogeneous equation $x'' + x = 0$. The solutions are $x_h(t) = a\cos t + b\sin t$.
2. Find any particular solution $x_p(t)$ to $x'' + x = 2$. By inspection, we see that $x_p(t) = 2$ works.
3. Add these together to get the general solution:

$$x(t) = x_h(t) + x_p(t) = a\cos t + b\sin t + 2.$$

Note that while the general solution above is unique, its presentation need not be.

For example, we could write it this way:

$$x(t) = x_h(t) + x_p(t) = a(2\cos t - 3\sin t) + b\sin t + (2 - \cos t + 8\sin t).$$

Here, the particular solution has (unnecessary) "extra terms" that vanish on the homogeneous part, $x'' + x = 0$.

The vanishing ideal and the model space

The function $f(x) = x_1(x_3 + 1)$ fits the following data:

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

To find the model space $\mathrm{Mod}(\mathcal{D}) = f + I(\mathcal{D})$, we need to find the vanishing ideal

$$I(\mathcal{D}) \subseteq R/I := \mathbb{F}[x_1, \dots, x_n]/\langle x_1^2 - 1, \dots, x_n^2 - n\rangle.$$

The polynomials that vanish on $\mathrm{s}_i = (s_{i1}, s_{i2}, s_{i3})$ is the ideal

$$\begin{aligned} I(\mathrm{s}_i) &= \big\{(x_1 - s_{i1})g_1(\mathrm{x}) + (x_2 - s_{i2})g_2(\mathrm{x}) + (x_3 - s_{i3})g_3(\mathrm{x}) \mid g_i(\mathrm{x}) \in R/I\big\} \\ &= \big\langle x_1 - s_{i1},\ x_2 - s_{i2},\ x_3 - s_{i3}\big\rangle. \end{aligned}$$

The vanishing ideal is thus

$$\begin{aligned} I(\mathcal{D}) &= I(\mathrm{s}_1) \cap I(\mathrm{s}_2) \cap I(\mathrm{s}_3) \\ &= \langle x_1 - 1, x_2 - 1, x_3 - 1\rangle \cap \langle x_1 - 1, x_2 - 1, x_3\rangle \cap \langle x_1, x_2, x_3\rangle. \end{aligned}$$

Note that this ideal has size $|I(\mathcal{D})| = |\,\mathrm{Mod}(\mathcal{D})| = 2^{8-3} = 32$. *(Why?)*

The vanishing ideal and the model space

The function $f(x) = x_1(x_3+1)$ fits the following data:

Input vectors:	$\mathbf{s}_1$	$\mathbf{s}_2$						$\mathbf{s}_3$
$x_1x_2x_3$	111	110	101	100	011	010	001	000
$f(x)$	0	1	?	?	?	?	?	0
Output values	t_1	t_2						t_3

We can compute the vanishing ideal in Macaulay2:

```
Q = ZZ/2[x1,x2,x3] / ideal(x1^2-x1, x2^2-x2, x3^2-x3);
I1 = ideal(x1-1, x2-1, x3-1);
I2 = ideal(x1-1, x2-1, x3);
I3 = ideal(x1, x2, x3-1);
I_D = intersect{I1,I2,I3};
```

The output is:

```
ideal(x1-x2, x2x3-x2-x3+1)
```

Thus, the model space consists of the 32 functions

$$\mathrm{Mod}(\mathcal{D}) = f + I(\mathcal{D}) = \{x_1(x_3+1) + (x_1+x_2)g_1 + (x_2x_3+x_2+x_3+1)g_2 \mid g_i \in R/I\}.$$

Inferring Boolean models

We just saw how to find the model space of a Boolean function $f\colon \mathbb{F}_2^n \to \mathbb{F}_n$.

To find the model space of a Boolean model $(f_1, \dots, f_n)$, we just do this for each coordinate.

Consider a set of data $\mathcal{D} = \big\{(\mathrm{s}_1, \mathrm{t}_1), \dots, (\mathrm{s}_k, \mathrm{t}_k)\big\}$, with

$$\text{Input vectors: } \mathrm{s}_1, \dots, \mathrm{s}_m \in \mathbb{F}^n$$

$$\text{Output vectors: } \mathrm{t}_1, \dots, \mathrm{t}_m \in \mathbb{F}^n$$

That is, $f(\mathrm{s}_i) = (f_1(\mathrm{s}_i), f_2(\mathrm{s}_i), \dots, f_n(\mathrm{s}_i)) = (t_{i1}, t_{i2}, \dots, t_{in}) = \mathrm{t}_i$.

We can encode this with n data sets of **input vectors** and *output values*:

$$\mathcal{D}_i = \big\{(\mathrm{s}_1, t_{1i}), (\mathrm{s}_2, t_{2i}), \dots, (\mathrm{s}_k, t_{1k})\big\}.$$

The model space of $\mathcal{D}$ is the direct product

$$\begin{aligned}\mathrm{Mod}(\mathcal{D}) &= \Big\{(f_1, \dots, f_n) \mid f_j(\mathrm{s}_i) = t_{ij} \text{ for all } i \text{ and } j\Big\} \\ &= \big[f_1 + I(\mathcal{D})\big] \times \cdots \times \big[f_n + I(\mathcal{D})\big] \\ &= \mathrm{Mod}(\mathcal{D}_1) \times \cdots \times \mathrm{Mod}(\mathcal{D}_n).\end{aligned}$$

An example

Consider the following model of the *lac* operon, which implicitly assumes that A degrades slower than M or B.

$$\begin{cases} f_M = x_A \\ f_B = x_M \\ f_A = L \vee (B \wedge L_m) \vee (A \wedge \overline{B}). \end{cases}$$

If lactose levels are low, then $L = L_m = 0$, and this model reduces to the following:

$$\begin{cases} f_1 = x_3 \\ f_2 = x_1 \\ f_3 = (x_2 + 1)x_3. \end{cases}$$

$$\begin{tikzcd} & & 011 \arrow[r] & 100 \arrow[dr] & & \\ 001 \arrow[r] & 101 \arrow[r] & 111 \arrow[r] & 110 \arrow[r] & 010 \arrow[r] & 000 \arrow[loop right] \end{tikzcd}$$

Let's find the model space of just the data given by the red nodes and edges.

```
R = ZZ/2[x1,x2,x3] / ideal(x1^2-x1, x2^2-x2, x3^2-x3)
I1 = ideal(x1, x2, x3-1);       I2 = ideal(x1-1, x2, x3-1);
I3 = ideal(x1-1, x2-1, x3-1);   I4 = ideal(x1-1, x2-1, x3);
I5 = ideal(x1, x2-1, x3);       I_D = intersect{I1,I2,I3,I4,I5};
```

The vanishing ideal consists of the 8 functions

$$I(\mathcal{D}) = \langle x_2x_3 + x_2 + x_3 + 1,\ x_1x_2 + x_1x_3 + x_1 + x_2 + x_3 + 1 \rangle,$$

and so the full model space is

$$\mathrm{Mod}(\mathcal{D}) = \big(f_1 + I(\mathcal{D}),\ f_2 + I(\mathcal{D}),\ f_3 + I(\mathcal{D})\big) = \big(x_3 + I(\mathcal{D}),\ x_1 + I(\mathcal{D}),\ (x_2+1)x_3 + I(\mathcal{D})\big).$$

An example (cont.)

Let's now suppose that we didn't *a priori* know a particular solution.

We'll use interpolation to find $f = (f_1, f_2, f_3)$ that fits the data. For example:

$$\begin{aligned} f_1(x) &= t_{11} r_1(x) + t_{21} r_2(x) + t_{31} r_3(x) + t_{41} r_4(x) + t_{51} r_5(x) \\ &= 1r_1(x) + 1r_2(x) + 1r_3(x) + 0r_4(x) + 0r_5(x) = r_1(x) + r_2(x) + r_3(x)\,, \end{aligned}$$

where

$$r_1(x) = \prod_{\substack{k=1 \\ k\neq 1}}^{5} (x_{\ell_k} - s_{k\ell_k}) = (x_{\ell_2} - s_{2\ell_2})(x_{\ell_3} - s_{3\ell_3})(x_{\ell_4} - s_{4\ell_4})(x_{\ell_5} - s_{5\ell_5}).$$

$$\begin{array}{c} s_1 = (0,0,1) \\ \downarrow \\ s_2 = (\underline{1},0,1) = t_1 \\ \downarrow \\ s_3 = (\underline{1},\underline{1},1) = t_2 \\ \downarrow \\ s_4 = (\underline{1},\underline{1},\underline{0}) = t_3 \\ \downarrow \\ s_5 = (0,\underline{1},\underline{0}) = t_4 \\ \downarrow \\ (0,0,\underline{0}) = t_5 \end{array}$$

Recall that ℓ_k is *any* coordinate in which s_1 differs from s_k.

skip $k = 1$

$$\begin{aligned} b_{12}(x) &= (x_1 - s_{21}) = x_1 + 1 \\ b_{13}(x) &= (x_1 - s_{31}) = x_1 + 1 \\ b_{14}(x) &= (x_1 - s_{41}) = x_1 + 1 \\ b_{15}(x) &= (x_2 - s_{52}) = x_2 + 1 \end{aligned}$$

Let's take $r_1(x) = (x_1 + 1)^3 (x_2 + 1) = (x_1 + 1)(x_2 + 1)$.

An example (cont.)

Recall that $b_{ik}(\mathrm{x}) = x_{\ell_k} - s_{k\ell_k}$, where ℓ_k is any coordinate that s_i differs from s_k.

$$
\begin{array}{c}
\mathrm{s}_1 = (0,0,1) \\
\downarrow \\
\mathrm{s}_2 = (1,0,1) = \mathrm{t}_1 \\
\downarrow \\
\mathrm{s}_3 = (1,1,1) = \mathrm{t}_2 \\
\downarrow \\
\mathrm{s}_4 = (1,1,0) = \mathrm{t}_3 \\
\downarrow \\
\mathrm{s}_5 = (0,1,0) = \mathrm{t}_4 \\
\downarrow \\
(0,0,0) = \mathrm{t}_5
\end{array}
$$

$$
\begin{aligned}
b_{21}(\mathrm{x}) &= (x_1 - s_{11}) = x_1 \\
&\text{skip } k = 2 \\
b_{23}(\mathrm{x}) &= (x_2 - s_{32}) = x_2 + 1 \\
b_{24}(\mathrm{x}) &= (x_2 - s_{42}) = x_2 + 1 \\
b_{25}(\mathrm{x}) &= (x_1 - s_{51}) = x_1
\end{aligned}
\qquad
\begin{aligned}
b_{31}(\mathrm{x}) &= (x_1 - s_{11}) = x_1 \\
b_{32}(\mathrm{x}) &= (x_2 - s_{22}) = x_2 \\
&\text{skip } k = 3 \\
b_{34}(\mathrm{x}) &= (x_3 - s_{43}) = x_3 \\
b_{35}(\mathrm{x}) &= (x_1 - s_{51}) = x_1
\end{aligned}
$$

$$
\begin{aligned}
b_{41}(\mathrm{x}) &= (x_1 - s_{11}) = x_1 \\
b_{42}(\mathrm{x}) &= (x_2 - s_{22}) = x_2 \\
b_{43}(\mathrm{x}) &= (x_3 - s_{33}) = x_3 + 1 \\
&\text{skip } k = 4 \\
b_{45}(\mathrm{x}) &= (x_1 - s_{51}) = x_1
\end{aligned}
\qquad
\begin{aligned}
b_{51}(\mathrm{x}) &= (x_2 - s_{12}) = x_2 \\
b_{52}(\mathrm{x}) &= (x_1 - s_{21}) = x_1 + 1 \\
b_{53}(\mathrm{x}) &= (x_1 - s_{31}) = x_1 + 1 \\
b_{54}(\mathrm{x}) &= (x_1 - s_{42}) = x_1 + 1 \\
&\text{skip } k = 5
\end{aligned}
$$

Recall that $x_i^k = x_i$, and $(x_j + 1)^k = x_j + 1$, so the "r-polynomials" are

$$
\begin{aligned}
r_1(\mathrm{x}) &= (x_1 + 1)(x_2 + 1) \\
r_2(\mathrm{x}) &= x_1(x_2 + 1) \\
r_3(\mathrm{x}) &= x_1 x_2 x_3 \\
r_4(\mathrm{x}) &= x_1 x_2 (x_3 + 1) \\
r_5(\mathrm{x}) &= (x_1 + 1) x_2
\end{aligned}
$$

An example (cont.)

We can now compute our particular solution (f_1, f_2, f_3) that fits the data, using:

$$\boxed{f_j(x) = t_{1j} r_1(x) + t_{2j} r_2(x) + \cdots + t_{mj} r_m(x)}.$$

$$\begin{array}{c} s_1 = (0,0,1) \\ \downarrow \\ s_2 = (1,0,1) = t_1 \\ \downarrow \\ s_3 = (1,1,1) = t_2 \\ \downarrow \\ s_4 = (1,1,0) = t_3 \\ \downarrow \\ s_5 = (0,1,0) = t_4 \\ \downarrow \\ (0,0,0) = t_5 \end{array}$$

$$\begin{aligned} f_1(x) &= t_{11} r_1(x) + t_{21} r_2(x) + t_{31} r_3(x) + t_{41} r_4(x) + t_{51} r_5(x) \\ &= r_1(x) + r_2(x) + r_3(x) \\ &= 1 + x_2 + x_1 x_2 x_3 \end{aligned}$$

$$\begin{aligned} f_2(x) &= t_{12} r_1(x) + t_{22} r_2(x) + t_{32} r_3(x) + t_{42} r_4(x) + t_{52} r_5(x) \\ &= r_2(x) + r_3(x) + r_4(x) \\ &= x_1 \end{aligned}$$

$$\begin{aligned} f_3(x) &= t_{13} r_1(x) + t_{23} r_2(x) + t_{33} r_3(x) + t_{43} r_4(x) + t_{53} r_5(x) \\ &= r_1(x) + r_2(x) \\ &= 1 + x_2. \end{aligned}$$

Our original model was $(f_1, f_2, f_3) = (x_3,\ x_1,\ x_3 + x_2 x_3)$, but our algorithm yielded

$$\begin{aligned} (f_1, f_2, f_3) &= (1 + x_2 + x_1 x_2 x_3,\ x_1,\ 1 + x_2) \\ &= (x_3,\ x_1,\ x_3 + x_2 x_3) + (1 + x_2 + x_3 + x_1 x_2 x_3,\ 0,\ 1 + x_2 + x_3 + x_2 x_3) \end{aligned}$$

Remark

Each polynomial in the 2nd term above is in the *vanishing ideal* I. (Why?)

An example (cont.)

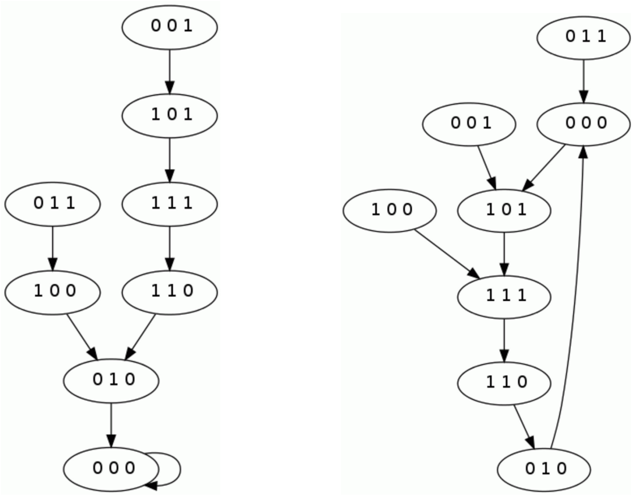


Figure: The original phase space (left), and the reverse-engineered phase space (right).

An example (cont.)

Now that we found a particular solution $f = (f_1, f_2, f_3)$ that fits the data, we need to (re)compute the ideal I of polynomials that vanish on the data.

```
R=ZZ/2[x1,x2,x3] / ideal(x1^2-x1, x2^2-x2, x3^2-x3);
```

$$s_1 = (0,0,1)$$
$$\downarrow$$
$$s_2 = (1,0,1) = t_1$$
$$\downarrow$$
$$s_3 = (1,1,1) = t_2$$
$$\downarrow$$
$$s_4 = (1,1,0) = t_3$$
$$\downarrow$$
$$s_5 = (0,1,0) = t_4$$
$$\downarrow$$
$$(0,0,0) = t_5$$

The ideal of polynomials that vanish on each s_k is:

```
I1 = ideal(x1, x2, x3-1);
I2 = ideal(x1-1, x2, x3-1);
I3 = ideal(x1-1, x2-1, x3-1);
I4 = ideal(x1-1, x2-1, x3);
I5 = ideal(x1, x2-1, x3);
```

The ideal of polynomials that vanish on *every* s_k is:

```
I = intersect{I1,I2,I3,I4,I5}
```

To compute a Gröbner basis:

```
G = gens gb I
```

The output is: `| x2x3+x2+x3+1 x1x2+x1x3+x1+x2+x3+1 |`

An example (cont.)

In conclusion, the set of all Boolean models that fit the data $\mathcal{D}$

$$001 \longrightarrow 101 \longrightarrow 111 \longrightarrow 110 \longrightarrow 010 \longrightarrow 000$$

i.e., the model space, is the set

$$F_1 \times F_2 \times F_3, \qquad F_j = f_j + I(\mathcal{D})$$

where $I(\mathcal{D})$ is the vanishing ideal

$$I(\mathcal{D}) = \langle g_1, g_2 \rangle = \big\langle 1 + x_2 + x_3 + x_2x_3,\ 1 + x_1 + x_2 + x_3 + x_1x_2 + x_1x_3 \big\rangle.$$

Our reverse-engineered BN is slighly different than the "true model":

$$\begin{aligned}(f_1, f_2, f_3) &= (1 + x_2 + x_1x_2x_3,\ x_1,\ 1 + x_2) \\ &= (x_3 + x_1g_1 + g_2,\ \ x_1,\ \ (x_2 + 1)x_3 + g_1)\end{aligned}$$

Note that $x_1g_1 + g_2$, 0, and g_1 *must* be in the vanishing ideal I.

An example (cont.)

Goal ("model selection")

We would like to recover functions in $F_j = f_j + I$ that have no "extra terms" in I.

$$001 \longrightarrow 101 \longrightarrow 111 \longrightarrow 110 \longrightarrow 010 \longrightarrow 000$$

For example, the following particular solution has "extra terms":

$$x'' + x = 2, \qquad x(t) = x_h(t) + x_p(t) = a\cos t + b\sin t + (2 + \underbrace{5\cos t - 4\sin t}_{\text{unnecessary; in } x_h(t)}).$$

One approach: the Gröbner normal form, which is the "remainder of f_j modulo I."

This does depends on the Gröbner basis, which depends on a choice of monomial ordering.

We can do this with Macaulay2, using the % symbol.

```
f1 = 1+x2+x1*x2*x3;
f2 = x1;
f3 = 1+x2;
f1%I; f2%I; f3%I;
(f1, f2, f3)
```

The output is: `(x3, x1, x2+1)`. Almost the original Boolean model!

Non-Boolean models

Just like the Boolean case, over a general finite field $\mathbb{F}_p$, it suffices to construct

$$r_i(\mathrm{x}) = \begin{cases} 1 & \mathrm{x} = \mathrm{s}_i \\ 0 & \mathrm{x} = \mathrm{s}_j,\ j \neq i \end{cases}$$

because then the following is a solution:

$$f(\mathrm{x}) = t_1 r_1(\mathrm{x}) + t_2 r_2(\mathrm{x}) + t_3 r_3(\mathrm{x}).$$

Over $\mathbb{F}_2$, our construction guaranteed $r_i(\mathrm{s}_i) \neq 0$, which is equivalent to $r_i(\mathrm{s}_i) = 1$.

Over $\mathbb{F}_p$, we have to be a little more careful. The following corrects for this:

$$r_i(\mathrm{x}) = \prod_{\substack{k=1 \\ k \neq i}}^{m} b_{ik}(\mathrm{x}), \qquad b_{ik}(\mathrm{x}) = \underbrace{(s_{i\ell_k} - s_{k\ell_k})^{p-2}}_{\text{ensures that } r_i(\mathrm{s}_i) = 1} (x_\ell - s_{k\ell_k})$$

An example over $\mathbb{F}_5$

Consider the following time series in a 3-node algebraic model over $\mathbb{F}_5$:

$$\mathrm{s}_1 = (2,0,0) \longrightarrow \mathrm{s}_2 = (4,3,1) = \mathrm{t}_1 \longrightarrow \mathrm{s}_3 = (3,1,4) = \mathrm{t}_2 \longrightarrow (0,4,3) = \mathrm{t}_3$$

$$r_i(\mathrm{x}) = \prod_{\substack{k=1 \\ k \neq i}}^{m} b_{ik}(\mathrm{x})$$

$$b_{ik}(\mathrm{x}) = (s_{i\ell_k} - s_{k\ell_k})^{p-2}(x_\ell - s_{k\ell_k})$$

Note that s_1 differs from s_2 and s_3 in the $\ell_k = 1$ coodinate, so this will work for each r_i.

Particularly useful identities are: $0 = 5$, $-1 = 4$, $-2 = 3$, $-3 = 2$, and $-4 = 1$.

Using our formulas for $b_{ij}(\mathrm{x})$, we compute:

$$\begin{aligned} b_{12}(\mathrm{x}) &= (s_{11} - s_{21})^3(x_1 - s_{21}) = (2-4)^3(x_1 - 4) = -8(x_1+1) = 2x_1 + 2 \\ b_{13}(\mathrm{x}) &= (s_{11} - s_{31})^3(x_1 - s_{31}) = (2-3)^3(x_1 - 3) = -x_1 + 3 = 4x_1 + 3\,. \end{aligned}$$

Therefore, the first r-polynomial is

$$r_1(\mathrm{x}) = b_{12}(\mathrm{x})b_{13}(\mathrm{x}) = (2x_1+2)(4x_1+3) = 8x_1^2 + 14x_1 + 6 = 3x_1^2 + 4x_1 + 1\,.$$

An example over $\mathbb{F}_5$ (cont.)

Similarlly, we can compute the other r-polynomials, and they are

$$\begin{aligned}
r_1(\mathrm{x}) &= b_{12}(\mathrm{x})b_{13}(\mathrm{x}) = (2x_1+2)(4x_1+3) = 8x_1^2+14x_1+6 = 3x_1^2+4x_1+1\\
r_2(\mathrm{x}) &= b_{21}(\mathrm{x})b_{23}(\mathrm{x}) = (3x_1+4)(x_1+2) = 3x_1^2+10x_1+8 = 3x_1^2+3\\
r_3(\mathrm{x}) &= b_{31}(\mathrm{x})b_{32}(\mathrm{x}) = (x_1+3)(4x_1+4) = 4x_1^2+16x_1+12 = 4x_1^2+x_1+2
\end{aligned}$$

Thus, the following functions fit the data:

$$\begin{aligned}
f_1(\mathrm{x}) &= t_{11}r_1(\mathrm{x}) + t_{21}r_2(\mathrm{x}) + t_{31}r_3(\mathrm{x})\\
&= 4(3x_1^2+4x_1+1) + 3(3x_1^2+3) + 0(4x_1^2+x_1+2)\\
&= x_1^2+x_1+3
\end{aligned}$$

$$\begin{aligned}
f_2(\mathrm{x}) &= t_{12}r_1(\mathrm{x}) + t_{22}r_2(\mathrm{x}) + t_{32}r_3(\mathrm{x})\\
&= 3(3x_1^2+4x_1+1) + 1(3x_1^2+3) + 4(4x_1^2+x_1+2)\\
&= 3x_1^2+x_1+4
\end{aligned}$$

$$\begin{aligned}
f_3(\mathrm{x}) &= t_{13}r_1(\mathrm{x}) + t_{23}r_2(\mathrm{x}) + t_{33}r_3(\mathrm{x})\\
&= 1(3x_1^2+4x_1+1) + 4(3x_1^2+3) + 3(4x_1^2+x_1+2)\\
&= 2x_1^2+2x_1+4
\end{aligned}$$

We have just found a single particular solution (f_1, f_2, f_3) that fits the data.

An example over $\mathbb{F}_5$ (cont.)

If $I(\mathrm{s}_i)$ is the ideal that vanishes on s_i, then the vanishing ideal $I(\mathcal{D})$ is

$$I(\mathcal{D}) = I(\mathrm{s}_1) \cap I(\mathrm{s}_2) \cap I(\mathrm{s}_3) \qquad \mathrm{s}_1 = (2,0,0)\,, \quad \mathrm{s}_2 = (4,3,1)\,, \quad \mathrm{s}_3 = (3,1,4)\,.$$

These are precisely the sets

$$I(\mathrm{s}_1) = \langle x_1 - 2, x_2, x_3 \rangle = \{(x_1-2)g_1(x) + x_2 g_2(x) + x_3 g_3(x)\}$$
$$I(\mathrm{s}_2) = \langle x_1 - 4, x_2 - 3, x_3 - 1 \rangle = \{(x_1-4)g_1(x) + (x_2-3)g_2(x) + (x_3-1)g_3(x)\}$$
$$I(\mathrm{s}_3) = \langle x_1 - 3, x_2 - 1, x_3 - 4 \rangle = \{(x_1-3)g_1(x) + (x_2-1)g_2(x) + (x_3-4)g_3(x)\}.$$

As before, we can compute this in Macaulay2:

```
R=ZZ/5[x1,x2,x3] / ideal(x1^5-x1, x2^5-x2, x3^5-x3);
I1 = ideal(x1-2, x2, x3);
I2 = ideal(x1-4, x2-3, x3-1);
I3 = ideal(x1-3, x2-1, x3-4);
I_D = intersect{I1,I2,I3};
gens gb I_D
```

A Gröbner basis for $I(\mathcal{D})$ is thus

$$\mathcal{G} = \{x_1 - 2x_2 - x_3 - 2,\ x_3^2 + 2x_2 - 2x_3,\ x_2x_3 + 2x_2 + x_3,\ x_2^2 + x_3\}.$$

An example over $\mathbb{F}_5$ (cont.)

We constructed three functions that fit the following data $\mathcal{D}$:

$$\mathrm{s}_1 = (2,0,0), \qquad \mathrm{s}_2 = (4,3,1) = \mathrm{t}_1, \qquad \mathrm{s}_3 = (3,1,4) = \mathrm{t}_2, \qquad \mathrm{t}_3 = (0,4,3).$$

Notice that the functions we found depend only on x_1. (*Why?*)

```
f1=x1*x1+x1+3;
f2=3x1^2+x1+4;
f3=2x1^2+2x1+4;
```

We can compute the Gröbner normal form in Macaulay2:

```
p1 = f1 % I_D;  p2 = f2 % I_D;  p3 = f3 % I_D;
```

The output is

$$(p_1, p_2, p_3) = (-x_3 - 1,\ x_2 - 2,\ -2x_3 + 1) = (4x_3 + 4,\ x_2 + 3,\ 3x_3 + 1).$$

The model space is thus

$$(4x_3 + 4,\ x_2 + 3,\ 3x_3 + 1) + I(\mathcal{D}) \times I(\mathcal{D}) \times I(\mathcal{D}),$$

where

$$I(\mathcal{D}) = \big\{(x_1 - 2x_2 - x_3 - 2)g_1 + (x_3^2 + 2x_2 - 2x_3)g_2 + (x_2x_3 + 2x_2 + x_3)g_3 + (x_2^2 + x_3)g_4 \mid g_i \in R/I\big\}.$$