PURM 2020: Generating Independent Data

Problem Statement Examples

Alan Ismaiel and Jason Shu

Penn Engineering

PRECISE
PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

Static Example

Consider 2 random variables A and B. A has 3 values, denoted as 1, 2 and 3. B has 2 values, denoted as red and blue.

$P(A = 1 \cap B = red)$	o00
$P(A = 1 \cap B = blue)$	o01
$P(A = 2 \cap B = red)$	o10
$P(A = 2 \cap B = blue)$	o11
$P(A = 3 \cap B = red)$	o20
$P(A = 3 \cap B = blue)$	o21

Static Example

We are given 1 = o00 + … + o21. 5 more equations are needed.

- A independent of B (provides 2 equations)
- P(B = red) = .3
- P(A = 1) = .6
- P(A = 2 AND B = blue) = .2

Static Example: Mathematica

Input Code

```
outcomes3 = {o00, o01, o10, o11, o20, o21}
eqs3 = {
Plus @@ outcomes3 == 1,
o00 == (o00 + o10 + o20) * (o00 + o01),
o10 == (o00 + o10 + o20) * (o10 + o11),
(o00 + o10 + o20) == .3,
(o00 + o01) == .6
o11 == .2
  };
constr3 = # \[Element] Interval[{0, 1}] & /@ outcomes3;
Solve[Join[eqs3, constr3], outcomes3, Reals]
```

Result

{{o00 -> {0.06}, o01 -> {0.14}, o10 -> {0.142857}, o11 -> {0.333333}, o20 -> {0.0971429}, o21 -> {0.226667}}}

The system is fully defined and data generation is possible

Time Invariant Example 1

Consider a random variable X with 3 potential values, denoted as 1, 2 and 3. Furthermore, define X such that it has 1 relevant time step.

$P(X_t = 1 \cap X_{t-1} = 1)$	o00
$P(X_t = 1 \cap X_{t-1} = 2)$	o01
$P(X_t = 1 \cap X_{t-1} = 3)$	o02
$P(X_t = 2 \cap X_{t-1} = 1)$	o10
$P(X_t = 2 \cap X_{t-1} = 2)$	o11
$P(X_t = 2 \cap X_{t-1} = 3)$	o12
$P(X_t = 3 \cap X_{t-1} = 1)$	o20
$P(X_t = 3 \cap X_{t-1} = 2)$	o21
$P(X_t = 3 \cap X_{t-1} = 3)$	o22

Time Invariant Example 1

Given:

- 1 = o00 + … + o22
- P(X_t = 1) = P(X_t-1 = 1)
- P(X_t = 2) = P(X_t-1 = 2)

- P(X_t = 3) = P(X_t-1 = 3) is given as well, but ultimately redundant due to other given information

Added (6 equations required):

- P(X_t-1 = 1) = .5
- P(X_t-1 = 2) = .25
- X_t is independent of X_t-1 = 1
  - Provides 2 equations
- P(X_t = 2 | X_t-1 = 2) = .4
- P(X_t = 3 | X_t-1 = 3) = .4

Time Invariant Example 1: Mathematica

Input Code

```
outcomes3 = {o00, o01, o02, o10, o11, o12, o20, o21, o22}

eqs3 = {

Plus @@ outcomes3 == 1,

o00 + o10 + o20 == o00 + o01 + o02,

o01 + o11 + o21 == o10 + o11 + o12,

o00 + o10 + o20 == .5,

o01 + o11 + o21 == .25,

o00 == (o00 + o01 + o02) * (o00 + o10 + o20),

o10 == (o10 + o11 + o12) * (o00 + o10 + o20),

o11 / (o01 + o11 + o21) == .4,

o22 / (o02 + o12 + o22) == .4

  };

constr3 = # \[Element] Interval[{0, 1}] & /@ outcomes3;

Solve[Join[eqs3, constr3], outcomes3, Reals]
```

Result

{{o00 -> {0.25}, o01 -> {0.125}, o02 -> {0.125}, o10 -> {0.125}, o11 -> {0.1}, o12 -> {0.025}, o20 -> {0.125}, o21 -> {0.025}, o22 -> {0.1}}}

The system is fully defined and data generation is possible

Time Invariant Example 2

Consider a single boolean variable T, defined with 2 relevant time steps

$P(T_t \cap T_{t-1} \cap T_{t-2})$	o000
$P(T_t \cap T_{t-1} \cap \neg T_{t-2})$	o001
$P(T_t \cap \neg T_{t-1} \cap T_{t-2})$	o010
$P(T_t \cap \neg T_{t-1} \cap \neg T_{t-2})$	o011
$P(\neg T_t \cap T_{t-1} \cap T_{t-2})$	o100
$P(\neg T_t \cap T_{t-1} \cap \neg T_{t-2})$	o101
$P(\neg T_t \cap \neg T_{t-1} \cap T_{t-2})$	o110
$P(\neg T_t \cap \neg T_{t-1} \cap \neg T_{t-2})$	o111

Time Invariant Example 2

We can assume the following information:

- 1 = o000 + … + o111

Stationary Assumption information:

- P(T_t) = P(T_t-1) = P(T_t-2)
- P(T_t AND T_t-1) = P(T_t-1 AND T_t-2)
- P(T_t AND not T_t-1) = P(T_t-1 AND not T_t-2)
- P(not T_t AND T_t-1) = P(not T_t-1 AND T_t-2)

It can be proven that by defining P(T_t AND T_t-1) = P(T_t-1 AND T_t-2), P(T_t AND not T_t-1) = P(T_t-1 AND not T_t-2), and P(not T_t AND T_t-1) = P(not T_t-1 AND T_t-2), then all the necessary equalities between time steps required for the stationary assumption will hold true. Thus, 4 more equations needed.

Time Invariant Example 2

Add the following information:

- P(T_t | T_t-1 and T_t-2) = .2
- P(T_t | not T_t-1 and not T_t-2) = .8
- P(T_t | T_t-1 and not T_t-2) = .5
- P(T_t | not T_t-1 and not T_t-2) = .5

Mathematica Output: {{o000 -> {0.0384615}, o001 -> {0.153846}, o010 -> {0.153846}, o011 -> {0.153846}, o100 -> {0.153846}, o101 -> {0.153846}, o110 -> {0.153846}, o111 -> {0.0384615}}}

Notes:

- The amount of relations that must hold for the stationary assumption increases exponentially as the number of relevant time steps increases
- We CANNOT conclude that only a single time step is relevant: the conditional expressions clearly show that both their results could impact the outcome
  - As such, the distribution of these variables currently break rules that would be required to be considered a Markov Chain

Time Invariant Example 2: Generation

From last slide:

{{o000 -> {0.0384615}, o001 -> {0.153846}, o010 -> {0.153846}, o011 -> {0.153846}, o100 -> {0.153846}, o101 -> {0.153846}, o110 -> {0.153846}, o111 -> {0.0384615}}}

Calculate conditional probabilities for current step given past two:

1. P(T_t | T_t-1 && T_t-2) = o000 / (o000 + o100) = .2
2. P(T_t | T_t-1 && not T_t-2) = o001 / (o001 + o101) = .5
3. P(T_t | not T_t-1 && T_t-2) = o010 / (o010 + o110) = .5
4. P(T_t | not T_t-1 && not T_t-2) = o011 / (o011 + o111) = .8

Suppose T_1 was False and T_2 was True. Generate T_3.

From Equation 3 above, P(T_3 | not T_2 && T_1) = .5; generate T_3 using a random number generator.

Suppose T_3 was True. Now generate T_4 using Equation 1, as T_2 and T_3 were True. Repeat generation.

Time Variant Example 1

Consider a system with 2 boolean variables, X and Y. X is defined with 1 relevant time step, Y is defined with no relevant time steps.

$P(Y_t \cap X_t \cap X_{t-1})$	o000
$P(Y_t \cap X_t \cap \neg X_{t-1})$	o001
$P(Y_t \cap \neg X_t \cap X_{t-1})$	o010
$P(Y_t \cap \neg X_t \cap \neg X_{t-1})$	o011
$P(\neg Y_t \cap X_t \cap X_{t-1})$	o100
$P(\neg Y_t \cap X_t \cap \neg X_{t-1})$	o101
$P(\neg Y_t \cap \neg X_t \cap X_{t-1})$	o110
$P(\neg Y_t \cap \neg X_t \cap \neg X_{t-1})$	o111

Time Variant Example 1

Given:

- 1 = o000 + … + o111
- P(X_t-1) = q1

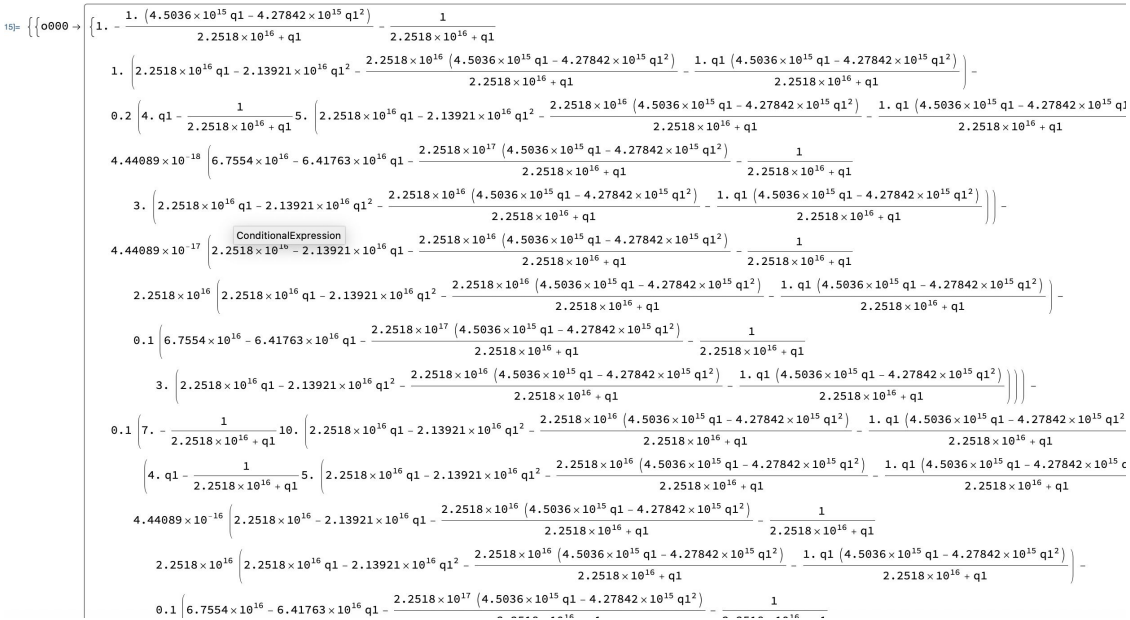
Added (6 equations required):

- X_t independent of X_t-1
- X_t independent of Y_t
- X_t independent of Y_t conditioned on X_t-1
- P(X_t) = P(X_t-1) * .95
- P(Y_t | X_t-1) = .2
- P(Y_t) = .3

Time Variant Example 1 Input

```
outcomes3 =
 Symbol[StringJoin["o", ToString@#]] & /@
  IntegerString[Range[0, 7], 2, 3]
eqs3 = {Plus @@ outcomes3 == 1,
o000 + o100 == (o000 + o001 + o100 + o101) * (o000 + o010 + o100 + o110),
o000 + o001 == (o000 + o001 + o100 + o101) * (o000 + o001 + o010 + o011),
o000 + o001 + o100 + o101 == .95 * (o000 + o010 + o100 + o110),
o000 + o010 + o100 + o110 == q1,
(o000 + o010) / (o000 + o010 + o100 + o110) == .2,
o000 + o001 + o010 + o011 == .3,
o000 / (o000 + o010 + o100 + o110) == ((o000 + o010) / (o000 + o010 + o100 + o110)) * ((o000 + o100) / (o000 + o010 + o100 + o110))
};
constr3 = # \[Element] Interval[{0, 1}] & /@ outcomes3;
Solve[Join[eqs3, constr3], outcomes3, Reals]
```

Time Variant Example 1 Output



Note: This is a small section of the output which defined all 8 parameters in terms of q1. Because all 8 parameters are fully defined, data generation is possible

Time Variant Example 2

Consider a single boolean variable T, defined with 2 relevant time steps

$P(T_t \cap T_{t-1} \cap T_{t-2})$	o000
$P(T_t \cap T_{t-1} \cap \neg T_{t-2})$	o001
$P(T_t \cap \neg T_{t-1} \cap T_{t-2})$	o010
$P(T_t \cap \neg T_{t-1} \cap \neg T_{t-2})$	o011
$P(\neg T_t \cap T_{t-1} \cap T_{t-2})$	o100
$P(\neg T_t \cap T_{t-1} \cap \neg T_{t-2})$	o101
$P(\neg T_t \cap \neg T_{t-1} \cap T_{t-2})$	o110
$P(\neg T_t \cap \neg T_{t-1} \cap \neg T_{t-2})$	o111

Time Variant Example 2

Additionally, consider the existence of 4 q parameters

$P(T_{t-1} \cap T_{t-2})$	q00
$P(T_{t-1} \cap \neg T_{t-2})$	q01
$P(\neg T_{t-1} \cap T_{t-2})$	q10
$P(\neg T_{t-1} \cap \neg T_{t-2})$	q11

These are defined similarly to the o parameters, albeit without any variables at time step t. We can assume that when generating data for any time t, we will already have the values of these 4 q parameters.

NOTE: Due to the given equation, we can put one q value in terms of the other 3. Therefore, only 3 of these values are necessary to define.

Time Variant Example 2

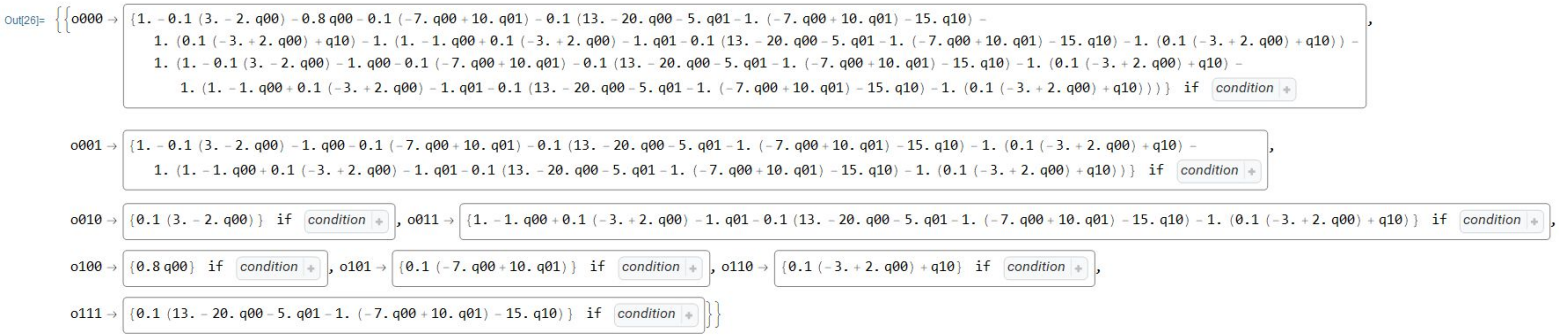
Consider 4 different specifications (Mathematica input format in blue):

- P(T_t) = .5*P(T_t-1) + .5*P(T_t-2)
  - o000 + o001 + o010 + o011 == .5 * (**q00** + **q01**) + .5 * (**q00** + **q10**)
- P(T_t and T_t-1) = .9 * P(T_t-1 and T_t-2)
  - o000 + o001 == .9 * **q00**
- P(T_t and T_t-2) = .3
  - o000 + o010 == .3
- P(T_t | T_t-1 AND T_t-2) = .2
  - o000 / (o000 + o001) == .2

This is in addition to the 4 specifications initially given (defining 3 of the **q parameters** in terms of o parameters, and the common given equation)

Time Variant Example 2

Output:



Q Conditions (system is conflicting if Qs fall outside these constraints):

```
(q00 > 0 && 0 < q01 < 0.190909 && q00 - 1.42857 q01 < 0 && 0.6 - 0.6 q00 - 1. q01 - q10 < 0 && -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0) ||
  (q00 > 0 && 0.190909 < q01 < 0.290244 && 0.6 - 0.6 q00 - 1. q01 - q10 < 0 &&
      -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0 && -0.75 + q00 + 2.5 q01 < 0) || (q00 > 0 && 0.290244 < q01 < 0.3 &&
      0.6 - 0.6 q00 - 1. q01 - q10 < 0 && -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0 && -0.75 + q00 + 2.5 q01 < 0) ||
  (q00 > 0 && 0.3 < q01 < 0.566667 && -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0 && 0.3 - 0.2 q00 - q10 < 0 && -0.85 + q00 + 1.5 q01 < 0) ||
  (0.190909 < q01 < 0.290244 && q00 - 1.42857 q01 < 0 && -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0 &&
      0.3 - 0.2 q00 - q10 < 0 && 0.75 - q00 - 2.5 q01 < 0) || (0.290244 < q01 < 0.3 && -0.866667 + 0.866667 q00 + 1. q01 + q10 < 0 &&
      0.3 - 0.2 q00 - q10 < 0 && -0.85 + q00 + 1.5 q01 < 0 && 0.75 - q00 - 2.5 q01 < 0)
```

Time Variant Example 2

STEP 1 (BASE CASE):

- q00 == .3
- q01 == .25
- q10 == .25

$$\{\{o000 \to \{0.06\}, o001 \to \{0.21\}, o010 \to \{0.24\}, o011 \to \{0.04\}, o100 \to \{0.24\}, o101 \to \{0.04\}, o110 \to \{0.01\}, o111 \to \{0.16\}\}\}$$

From here, deduce the next q values as follows:

- q00 == o000 + o001 == .27
- q01 == o010 + o011 == .28
- q10 == o100 + o101 == .28

$$\{\{o000 \to \{0.054\}, o001 \to \{0.189\}, o010 \to \{0.246\}, o011 \to \{0.061\}, o100 \to \{0.216\}, o101 \to \{0.091\}, o110 \to \{0.034\}, o111 \to \{0.109\}\}\}$$

STEP 2:

- Replace the previous q values with the ones deduced from step 1

From here, deduce the next q values as follows:

- q00 == o000 + o001 == .243
- q01 == o010 + o011 == .307
- q10 == o100 + o101 == .307

$$\{\{o000 \to \{0.0486\}, o001 \to \{0.1701\}, o010 \to \{0.2514\}, o011 \to \{0.0799\}, o100 \to \{0.1944\}, o101 \to \{0.1369\}, o110 \to \{0.0556\}, o111 \to \{0.0631\}\}\}$$

STEP 3:

- Replace the previous q values with the ones deduced from step 2

And so on…

Time Variant Example 2: Notes

It took some time determining a valid specification in terms of the q parameters, but this difficulty is inherent in every problem we've come across.

Particular to time variant examples is the difficulty in specifying valid q parameters, not just to generate a single step, but to generate new valid q parameters that can go on to specify more steps without inconsistency. Dealing with these conflicts will likely require more considered user input equations allowing for a wide range of possible q values.