Querying Bayesian model output with PostgreSQL

Wellington PostgreSQL Users Group

Finlay Thompson

17 September 2015

DRAGONFLY
Data Science

Bayesian modelling

Bayesian analysis involves converting data and conceptual models into probability distributions.

We interpret probabilities in the broad sense that:

- a probability p is a number between 0 and 1
- where $p = 1$ corresponds to TRUE
- and $p = 0$ corresponds to FALSE
- probabilities measure our confidence in a statement

Note that traditional 20th century statistics understands probabilities as the ratios coming from repeated experiments (think coin tosses).

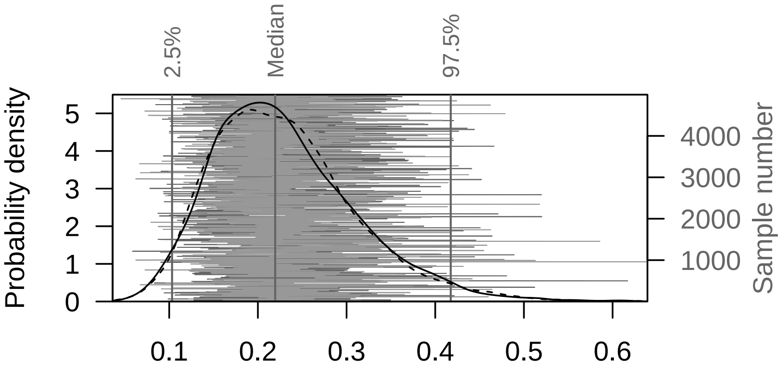
$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$

D represents the data, and is typically records in a database.

θ represents the parameters of some kind of model.

$P(D|\theta)$ is the *likelihood* distribution. Probability of measuring D given θ.

$P(\theta|D)$ is the *posterior* distribution. It represents the result of fitting the data D to the model θ.



DRAGONFLY
Data Science

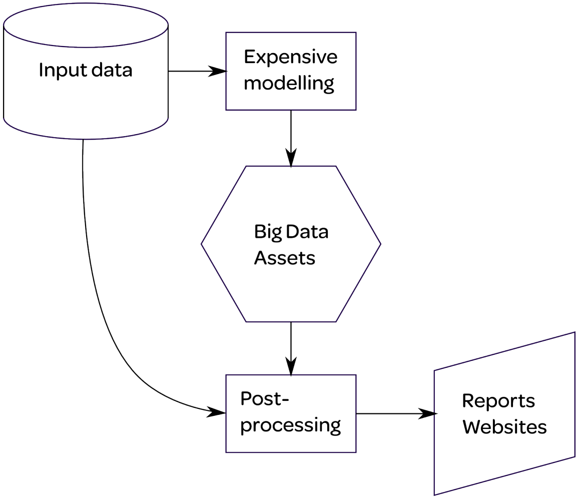
We use Monte Carlo Markov chain (MCMC) methods that are slow, and produce lots of output.

MCMC methods are **accurate**.

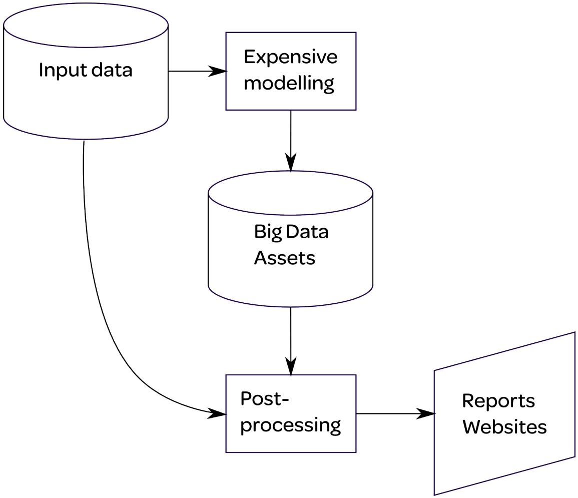
The posterior output is in the form of *samples from the posterior*. In practice this means 4000 values per parameter.

Typically models have hundreds of parameters.

DRAGONFLY
Data Science



DRAGONFLY
Data Science



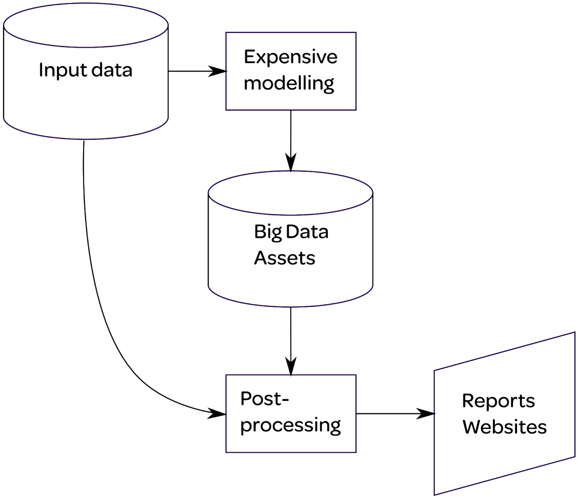
DRAGONFLY
Data Science

Another advantage of MCMC methods is **flexibility**.

We use the parameter samples to calculate samples for every desired output.

For example, we can produce 4000 samples for each of approx 1.5 million fishing events.

This output is big and expensive. So we moved it into PostgreSQL.



DRAGONFLY
Data Science

Storing and querying distributions

DRAGONFLY
Data Science

The estimates are stored in the form of arrays of integers.

They represent *uncertain* quantities, with each value in the array a realisation from the (unknown) posterior distribution.

The order of the arrays are significant, preserving the correlation structure of the estimates.

```sql
CREATE TABLE estimate (
    model_id    INTEGER REFERENCES model(id),
    effort_id   INTEGER NOT NULL,
    observed    INTEGER, -- null if effort not observed
    estimate    INTEGER[] NOT NULL
);
```

DRAGONFLY
Data Science

This results in storing a large quantity of data.

Currently around 12 GB

Need to aggregate the data, so integer array sums!

The standard function for aggregating integer arrays is a bit slow.

It checks the lengths of the arrays, and checks the types of arguments, checks checks checks.

I took the library function and ripped the checking out!

Open source for the win!

```c
#include "postgres.h"
#include "fmgr.h"
#include "utils/array.h"

#ifdef PG_MODULE_MAGIC
PG_MODULE_MAGIC;
#endif

Datum int_array_sum(PG_FUNCTION_ARGS);

PG_FUNCTION_INFO_V1(int_array_sum);
Datum
int_array_sum(PG_FUNCTION_ARGS) {

    ArrayType * state = PG_GETARG_ARRAYTYPE_P(0);
    ArrayType * new = PG_GETARG_ARRAYTYPE_P(1);

    int numargs = ARR_DIMS(state)[0];
    int * state_ptr = (int *) ARR_DATA_PTR(state);
    int * new_ptr = (int *) ARR_DATA_PTR(new);

    int i;
    for (i = 0; i < numargs; i++)
        state_ptr[i] += new_ptr[i];


    PG_RETURN_ARRAYTYPE_P(state);
}
```

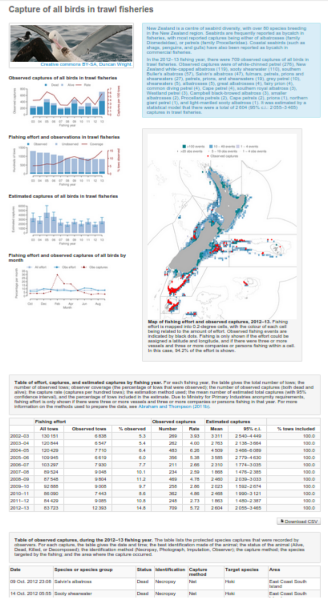
```sql
SET search_path = public;

CREATE OR REPLACE FUNCTION int_array_sum(int[], int[])
RETURNS int[]
AS '$libdir/intarraysum', 'int_array_sum'
LANGUAGE C IMMUTABLE STRICT;


DROP AGGREGATE IF EXISTS sum(int[]);
CREATE AGGREGATE sum (int[]) (
        SFUNC = int_array_sum,
        STYPE = int[]
);
```

Result is queries that are in the order of 1000 times faster.

The result is much more flexibility in reporting, and happy clients.



Capture of all birds in trawl fisheries

Creative commons BY-SA, Duncan Wright.

New Zealand is a centre of seabird diversity, with over 80 species breeding in the New Zealand region. Seabirds are frequently reported as bycatch in fisheries, with most reported captures being either of albatrosses (family Diomedeidae), or petrels (family Procellariidae). Coastal seabirds (such as shags, penguins, and gulls) have also been reported as bycatch in commercial fisheries.

In the 2012–13 fishing year, there were 709 observed captures of all birds in trawl fisheries. Observed captures were of white-chinned petrel (276), New Zealand white-capped albatross (119), sooty shearwater (110), southern Buller's albatross (57), Salvin's albatross (47), fulmars, petrels, prions and shearwaters (27), petrels, prions, and shearwaters (19), grey petrel (10), shearwaters (5), albatrosses (5), great albatrosses (4), fairy prion (4), common diving petrel (4), Cape petrel (4), southern royal albatross (3), Westland petrel (3), Campbell black-browed albatross (3), smaller albatrosses (2), Procellaria petrels (2), Cape petrels (2), prions (1), northern giant petrel (1), and light-mantled sooty albatross (1). It was estimated by a statistical model that there were a total of 2 604 (95% c.i.: 2 055–3 465) captures in trawl fisheries.

Observed captures of all birds in trawl fisheries

Fishing effort and observations in trawl fisheries

Estimated captures of all birds in trawl fisheries

Fishing effort and observed captures of all birds by month

Map of fishing effort and observed captures, 2012–13. Fishing effort is mapped into 0.2-degree cells, with the colour of each cell being related to the amount of effort. Observed fishing events are indicated by black dots. Fishing is only shown if the effort could be assigned a latitude and longitude, and if there were three or more vessels and three or more companies or persons fishing within a cell. In this case, 94.2% of the effort is shown.

Table of effort, captures, and estimated captures by fishing year. For each fishing year, the table gives the total number of tows; the number of observed tows; observer coverage (the percentage of tows that were observed); the number of observed captures (both dead and alive); the capture rate (captures per hundred tows); the estimation method used; the mean number of estimated total captures (with 95% confidence interval), and the percentage of tows included in the estimate. Due to Ministry for Primary Industries anonymity requirements, fishing effort is only shown if there were three or more vessels and three or more companies or persons fishing in that year. For more information on the methods used to prepare the data, see Abraham and Thompson (2011b).

	Fishing effort			Observed captures		Estimated captures		
	All tows	Observed tows	% observed	Number	Rate	Mean	95% c.i.	% tows included
2002–03	130 151	6 838	5.3	269	3.93	3 311	2 540–4 449	100.0
2003–04	120 844	6 547	5.4	262	4.00	2 763	2 138–3 664	100.0
2004–05	120 429	7 710	6.4	483	6.26	4 509	3 466–6 089	100.0
2005–06	109 945	6 619	6.0	356	5.38	3 585	2 779–4 630	100.0
2006–07	103 297	7 930	7.7	211	2.66	2 310	1 774–3 035	100.0
2007–08	89 524	9 048	10.1	234	2.59	1 868	1 476–2 385	100.0
2008–09	87 548	9 804	11.2	469	4.78	2 460	2 039–3 033	100.0
2009–10	92 888	9 008	9.7	258	2.86	2 023	1 592–2 674	100.0
2010–11	86 090	7 443	8.6	362	4.86	2 468	1 990–3 121	100.0
2011–12	84 429	9 085	10.8	248	2.73	1 863	1 480–2 387	100.0
2012–13	83 723	12 393	14.8	709	5.72	2 604	2 055–3 465	100.0

Download CSV

Table of observed captures, during the 2012–13 fishing year. The table lists the protected species captures that were recorded by observers. For each capture, the table gives the date and time; the best identification made of the animal; the status of the animal (Alive, Dead, Killed, or Decomposed); the identification method (Necropsy, Photograph, Imputation, Observer); the capture method; the species targeted by the fishing; and the area where the capture occurred.

Date	Species or species group	Status	Identification	Capture method	Target species	Area
09 Oct. 2012 23:08	Salvin's albatross	Dead	Necropsy	Net	Hoki	East Coast South Island
14 Oct. 2012 05:55	Sooty shearwater	Dead	Necropsy	Net	Hoki	East Coast South

Reporting is possible in many different slices:

- Fishing year
- target species of fishers
- reporting areas
- type of vessel

Each of these estimates has an accurate estimate of uncertainty, with published confidence intervals.

For a publicly visible example, see `https://data.dragonfly.co.nz/psc/`.