Analysis of birth cohort studies in BayesDB

Leonardo Casarsa, Belhal Karimi and Vikash K. Mansinghka

Abstract

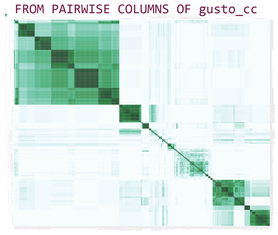
The Bill & Melinda Gates Foundation has a multi-year project on knowledge integration for Healthy Birth, Growth, and Development (HBGDki). The core aim is to make it possible to perform empirically grounded policy design, advocacy, and field treatment around growth insults such as stunting due to malnutrition. Tactics include gathering data on >5M children; funding new large-scale studies that link prenatal maternal & child health with family conditions and cognitive outcomes; and performing new kinds of analysis of existing data sources. Probabilistic programming plays an important role, addressing two key challenges: (i) there is far more routine data cleaning, exploration, imputation, and predictive modeling work than can be practically performed by the >20 person quantitative staff and (ii) new data sources such as GUSTO and INTERGROWTH present methodological challenges (around high dimensions, heterogeneous sources, etc) that currently lack solutions

This poster presents preliminary data sketches obtained by short MML and BQL programs that BMGF viewed as core validation that BayesDB and probabilistic programming can automate routine baseline modeling and data exploration, and produces results that are in accord with common-sense knowledge. It also shows preliminary evidence for predictive signals --- between micronutrients in the mother's blood at week 26 of pregnancy; anthropometry at birth; and cognitive outcomes at month 24, as assessed by Bayley scores --- that may warrant further in-depth exploration.

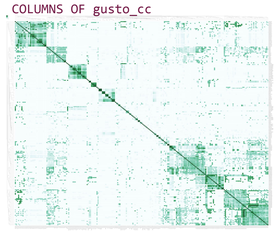
Dependence Probability

A key task in exploratory data analysis is to identify the variables in the database that are predictive of each other. In BayesDB, the evidence for predictive relationship is quantified as the dependence probability, i.e., the probability that the mutual information between two variables is nonzero (**figure**: left). An alternative measure is the significant ($p < 0.05$) correlation between pairwise variables (**figure**: right). Since the results from correlation are noisier, the right heatmap is considerably sparser.

```
ESTIMATE DEPENDENCE PROBABILITY
FROM PAIRWISE COLUMNS OF gusto_cc
```



```
ESTIMATE CORRELATION FROM PAIRWISE
COLUMNS OF gusto_cc
```



Acknowledgements & References

The authors would like to thank Feras Saad, Gregory Marton and Taylor Campbell for helpful feedback and discussions. This research was supported by DARPA (PPAML program, contract number FA8750-14-2-0004), IARPA (under research contract 2015-15061000003), the Office of Naval Research (under research contract N000141310333), the Army Research Office (under agreement number W911NF-13-1-0212), and gifts from Analog Devices and Google.

[1] V. Mansinghka, R. Tibbetts, J. Baxter, P. Shafto & B. Eaves (2015). BayesDB: A probabilistic programming system for querying the probable implications of data. arXiv preprint arXiv 1512.05006.

GUSTO - Growing Up in Singapore Towards healthy Outcomes

Growing Up in Singapore Towards healthy Outcomes (GUSTO) is the largest birth cohort study in Singapore yet, aimed at studying how a mother's diet and lifestyle during pregnancy impact the growth of her children.

We trained 32 models for 100 iterations on a subset of the GUSTO datatable using Crosscat.
Hierarchical clustering of the dependence probability grouped together variables that:

- Had same measurement time (e.g., at Pregnancy Week 26)
- Belonged to Father or Mother
- Regarded anthropometric measurements of the toddlers
- Regarded psychological scores (Bayley or Beck) of the toddlers

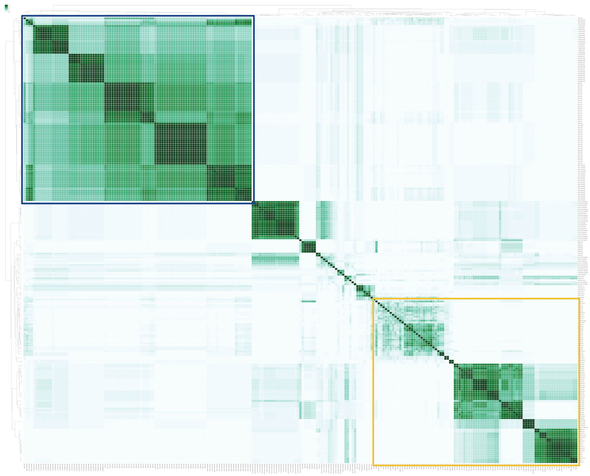
Variable name	Description
BDI01	Beck Depression Inventory-Sadness
BDI02	Beck Depression Inventory-Pessimism
BDI03	Beck Depression Inventory-Past Failure
BDI04	Beck Dep Inv-Loss of Pleasure
BDI05	Beck Dep Inv-Guilty Feelings
BDI06	Beck Dep Inv-Punishment Feelings
BDI07	Beck Depression Inventory-Self-Dislike
BDI08	Beck Dep Inv-Self-Criticalness
BDI09	Beck Dep Inv-Suicidal Thoughts/Wishes
BDI10	Beck Depression Inventory-Crying
BDI11	Beck Depression Inventory-Agitation
BDI12	Beck Dep Inv-Loss of Interest
BDI13	Beck Depression Inventory-Indecisiveness
BDI14	Beck Depression Inventory-Worthlessness
BDI15	Beck Depression Inventory-Loss of Energy
BDI16	Beck Dep Inv-ChangesSleep Pattern
BDI17	Beck Depression Inventory-Irritability
BDI18	Beck Dep Inv-Changes Appetite
BDI19	Beck Dep Inv-Concentration Difficulty
BDI20	Beck Dep Inv-Tiredness or Fatigue
BDI21	Beck Dep Inv-Loss of Interest in Sex
BDITOT	Beck Depression Inventory-Total Score

Variable name	Description
F_SYNEXT_M48	Syndr Externlz Prob Raw Tot Eval=FATH
F_SYNATT_M48	Syndr Attention Prob Raw Tot Eval=FATH
F_SYNAGG_M48	Syndr Aggr Behav Raw Tot Eval=FATH
F_DSMADH_M48	DSM AtnDeficit/Hyper Prob RawTot Eval=FATH
F_DSMODP_M48	DSM Oppos Def Prob Raw Tot Eval=FATH
F_SYNINT_M48	Syndr Intrnlz Prob Raw Tot Eval=FATH
F_SYNWDR_M48	Syndr Withdrawn Raw Tot Eval=FATH
F_SYNSOC_M48	Syndr Somatic Compl Raw Tot Eval=FATH
F_SYNSLP_M48	Syndr Sleep Prob Raw Tot Eval=FATH
F_SYNEMR_M48	Syndr Emotion Reac Raw Tot Eval=FATH
F_SYNADP_M48	Syndr Anxious/Depr Raw Tot Eval=FATH
F_SDSMSU_M48	Synd and DSM Total Score Sum Eval=FATH
F_DSMPDV_M48	DSM Pervasiv Dev Prob Raw Tot Eval=FATH
F_DSMAFF_M48	DSM Affective Prob Raw Tot Eval=FATH
F_DSMANX_M48	DSM Anxiety Prob Raw Tot Eval=FATH

Variable name	Description
M_SYNEXT_M48	Syndr Externlz Prob Raw Tot Eval=MOTH
M_SYNATT_M48	Syndr Attention Prob Raw Tot Eval=MOTH
M_SYNAGG_M48	Syndr Aggr Behav Raw Tot Eval=MOTH
M_DSMADH_M48	DSM AtnDeficit/Hyper Prob RawTot Eval=MOTH
M_DSMODP_M48	DSM Oppos Defian Prob Raw Tot Eval=MOTH
M_SYNINT_M48	Syndr Intrnlz Prob Raw Tot Eval=MOTH
M_SYNWDR_M48	Syndr Withdrawn Raw Tot Eval=MOTH
M_SYNSOC_M48	Syndr Somatic Compl Raw Tot Eval=MOTH
M_SYNSLP_M48	Syndr Sleep Prob Raw Tot Eval=MOTH
M_SYNEMR_M48	Syndr Emotion Reac Raw Tot Eval=MOTH
M_SYNADP_M48	Syndr Anxious/Depr Raw Tot Eval=MOTH
M_SDSMSU_M48	Synd and DSM Total Score Sum Eval=MOTH
M_DSMPDV_M48	DSM Pervasiv Dev Prob Raw Tot Eval=MOTH
M_DSMAFF_M48	DSM Affective Prob Raw Tot Eval=MOTH
M_DSMANX_M48	DSM Anxiety Prob Raw Tot Eval=MOTH

Variable name	Description
BDI02_PREGW26	Beck Depression Inventory-Pessimism
BDI16_PREGW26	Beck Dep Inv-ChangesSleep Pattern
BDI21_PREGW26	Beck Dep Inv-Loss of Interest in Sex

```
CREATE GENERATOR gusto_cc FOR gusto_table
   USING crosscat (BMI_DELIV NUMERICAL, COUNTRY CATEGORICAL, GUESS(*))
INITIALIZE 32 MODELS FOR gusto_cc
ANALYZE gusto_cc FOR 100 ITERATIONS
ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF gusto_cc
```



Variable name	Description
BALYMOR	Bayley Motor - Raw Score
BALYSER	Bayley Social-Emotional - Raw
BALYCUR	Bayley Community Use - Raw Score
BALYFAR	Bayley Functional Pre-Academic - Raw
BALYSOR	Bayley Social - Raw
BALYCOMR	Bayley Communication - Raw Score
BALYCNCC	Bayley Conceptual - Composite
BALYCNCS	Bayley Conceptual - Scaled
BALYHSR	Bayley Health and Saftey - Raw Score
BALYSCR	Bayley Self Care - Raw Score
BALYPRAS	Bayley Practical - Scaled
BALYPRAC	Bayley Practical - Compsite
BALYABPR	Bayley GAP - percentile rank
BALYHLR	Bayley Home Living - Raw Score
BALYSOPR	Bayley Social Composite percentile rank
BALYLSR	Bayley Leisure Behavior - Raw Score
BALYSOCC	Bayley Social Composite - Raw Score
BALYSOCS	Bayley Social Composite - Scaled
BALYSOR	Bayley Social - Raw
BALYLGS	Bayley Language - Scaled
BALYSCS	Bayley Self Care - Scaled
BALYHLS	Bayley Home Living - Scaled
BALYHSS	Bayley Health and Saftey - Scaled
BALYMOS	Bayley Motor - Scaled Score
BALYSOS	Bayley Social - Scaled
BALYFAS	Bayley Functional Pre-Academic - Scaled
BALYCUS	Bayley Community Use - Scaled
BALYCOMS	Bayley Communication - Scaled Score
BALYSEPR	Bayley Social-Emotion - Percentile rank
BALYCGPR	Bayley Cognitive - Percentile rank
BALYMTPR	Bayley Motor - Percentile rank

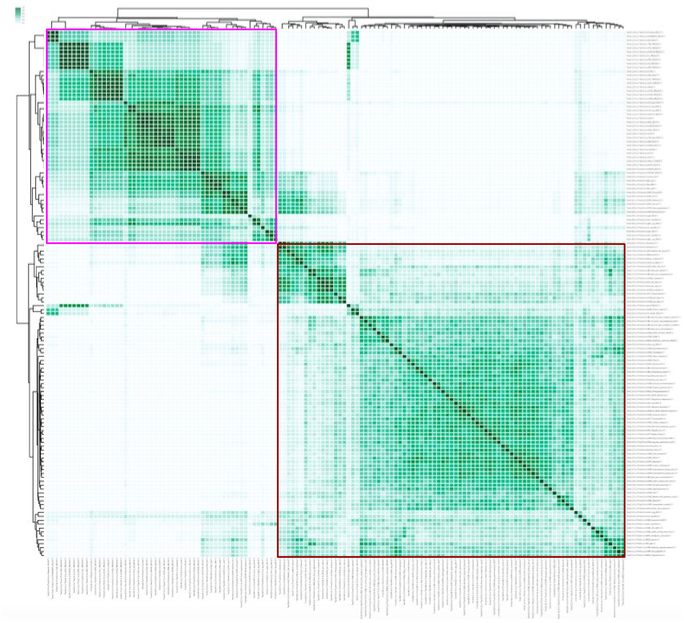
Variable name	Description
M_ZN	Zinc
M_MG	Magnesium
M_FE	Iron
M_CU	Copper
M_FERRITIN	Ferritin
SSFTMM	Subscapular skinfold thickness (mm)
TSFTMM	Triceps skinfold thickness (mm)
GAGEHX	Gest age at end of pregnancy (days)
GAGEBRTH	Gestational age at birth (days)
GAGEDAYS	Gestational age at examination (days)
BIRTHWT	Birth weight (gm)
HAZ	Length/height for age z-score
LENCM	Recumbent length (cm)
HCAZ	Head circum for age z-score
HCIRCM	Head circumference (cm)
WTKG	Weight (kg)
WAZ	Weight for age z-score
BIRTHWT	Birth weight (gm)
MUACCM	Mid upper-arm circumference (cm)
ABCIRCM	Abdominal circumference (cm)
WHZ	Weight for length/height z-score
BAZ	BMI for age z-score
BMI	BMI (kg/m**2)

Intergrowth - Interbio

INTERGROWTH-21st is a population-based multiethnic cohort aimed at developing International standards for fetal and postnatal Growth based on a prescriptive approach.
The study is taking place in eight nations, namely Brazil, China, India, Italy, Kenya, Oman, UK and USA, and it focus on collecting information regarding growth, health and nutrition from early pregnancy to infancy in healthy growth environments.

The INTERBIO-21st Study is an extension to the INTERGROWTH-21st project, aimed at improving characterization from preterm birth syndromes and intrauterine growth restriction. The data gathered encompasses standardised information on nutrition, pregnancy outcomes, newborn anthropometric measurements, and on growth and development up to 48 months of age.

```
CREATE GENERATOR intergrowth_cc FOR intergrowth_table
   USING crosscat (SEX CATEGORICAL, Alpha Numerical, GUESS(*))
INITIALIZE 32 MODELS FOR intergrowth_cc
ANALYZE intergrowth_cc FOR 100 ITERATIONS
ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF intergrowth_cc
```



Variable name	Description
PRETERM	Preterm Birth
GAGEBRTH	Gestational age at birth (days)
TAD	Teesside anemia in diabetes
OFD	Oral facial digital syndrome
BPD	Borderline personality disorder
APD	Auditory processing disorders
SEX	Sex
Z.AC	Pepcid AC disease at z score
Alpha	Growth coeff
A.OFD	Oral facial digital syndrome at z score
Z.BPD	Borderline personality disorder at z score
Z.LN	Lupus nephritis
LENCM	Recumbent length (cm)
WAZ	Weight for age z-score
BMI	BMI (kg/m**2)
WTKG	Weight (kg)
Linf	Growth coeff
SGA	Substantial gainful activity
Country	Country of Study
Mat_hgt	Maternal height
Mat_BMI	Maternal bmi
Mat_wgt	Maternal weight
SUPP_Vitamin_D	Supplement of Vitamin D
Multivitamins	Multivitamins

Variable name	Description
Betelnut	Use of Betelnut
Mode_of_delivery	Mode of delivery
Onset_of_labour	Beginning of the labor
Num_prev_births	Number of previous births
Num_prev_pregnancies	Number of previous pregnancies
Nuts_per_day	Nuts eaten per day
Sniffed_per_day	Sniffed per day

Variable name	Description
Yrs_since_last_pregnancy	Years since last pregnancy
MEDS_Antibiotics_antivirals	Antibiotics taken
OBS_Choriocarcinoma	Choriocarcinoma observed
Cig_per_day	Cigarettes smokes per day
Positiv_syphilis	Positive to Syphilis
Extrauterine_ectopic	Gestatio is extrauterine in nature
Alcohol_units.x	Unit of alcohols consumed by male
Supp_Selenium	Selenium supplements
MEDS_Aspirin	Aspirin consumed
COND_HIV_AIDS.x	HIV positivity within males