Probabilistic programming

Vid Stropnik, 63200434

Introduction

We are given data about the amount of resources 50 successful startups have invested in different sectors. Based on this information, we want to understand the most prosperous investment strategy. Furthermore, our goal is to understand potential differences in spending between the three states, wherein the data was sampled. All conclusions were achieved using the *Stan* probabilistic programming language and the work is reproducible and available on Github.

I. Setup

The dependency that we assume between our variables for this model is formalized in Equation 1;

$$\mathcal{P} = \beta_r * \mathcal{R} + \beta_m * \mathcal{M} + \beta_a * \mathcal{A} \tag{1}$$

where $\mathcal{P}, \mathcal{R}, \mathcal{M}, \mathcal{A}$ denote the profit, the research, marketing and administration spends respectively. The relation contains no intercept under the premise that no investments in any sector should correspond to a null profit. Another assumption of our model is that the observed profits are independent and homoscedastic around some mean. Due to the Central Limit Theorm, we assume that they are sampled from the normal distribtuion. We also put normal priors with sufficiently large variances ($N(0, 20)$) on our parameters, while the longer tails of the Cauchy prior are used for the deviation.

II. Composite relations of Investment sectors

The results of the MCMC process are shown in Table I. As a sanity check, we also examine the goodness of fit of our derived linear dependence in Figure 1. We notice that the weight (and line slope) of the research spend is distributed around the largest of means μ. By comparing the MCMC samples, we can formally express the truth stated in Eq. 2;

$$P(\beta_r > \beta_a > \beta_m | \mathcal{P}) = 0.96. \tag{2}$$

By using the weight estimates, we come to the sector spend proportional conclusions, logged in the rightmost column in Table I. As the observed data highly favor high-spending in marketing, we shouldn't infer a per-dollar proportional sharing strategy. Let these weights rather be interpreted that when a funding opportunity arises, we should prioritize the research sector 64%, the admin 29% and the marketing 7% of the time.

TABLE I
PARAMETERS OF POSTERIOR NORMAL DISTRIBUTIONS OF WEIGHTS β.

Value	Posterior mean μ	Posterior st.d. σ	$\mu(\beta)$ proportion[%]
β_r	0.7186	$16 \cdot e^{-4}$	63.67
β_m	0.0816	$5 \cdot e^{-4}$	29.10
β_a	0.3284	$7 \cdot e^{-4}$	7.23

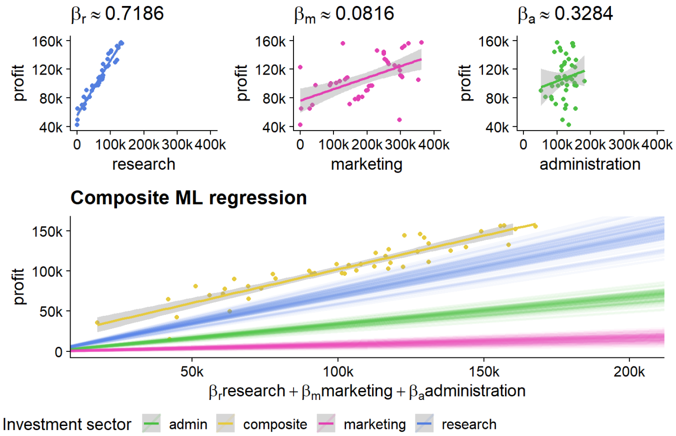

Fig. 1. Linear models fitted three investment sectors. The bottom figure shows the model (in yellow) fitted in composite space, calculated by our probabilistic programming approach. The blue, green and pink beams each correspond to lines, governed by the 100 beta samples from derived final distributions.

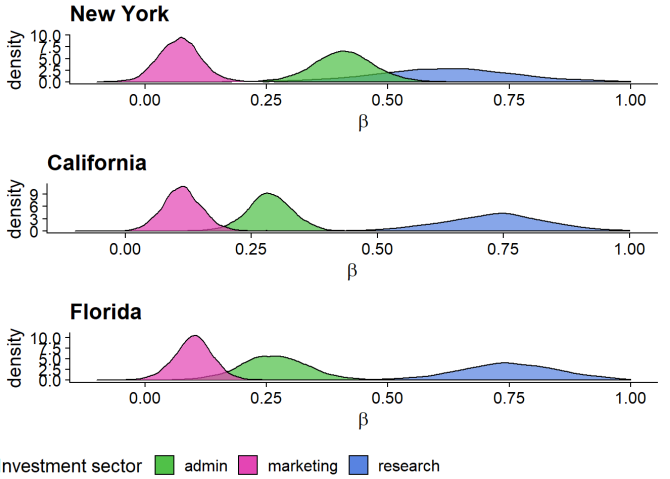

Fig. 2. The distributions of weights models. The three subplots correspond to different states where the observed businesses are located.

III. State differences

The differences in β values between the three states are shown in Figure 2. In it, we can notice that the spend in the administrative sector was more instrumental in New York, when compared to California and Florida - with a $\sim 12.85\%$ chance of being more important than research spend. The investment strategies are summarized in the table below:

State	NY	Cali	Florida
Research [%]	56	65	67
Admin [%]	37	25	24
Marketing [%]	7	10	9