The BRAIM Model

Attila Gulyás-Kovács

October 10, 2016

1 Description of BRAIM

The Bayesian regression allelic imbalance model (BRAIM) [1] assumes that genes are imprinted independently of each other[1]. The observed variable for mouse individual i and gene g is Y_{ig}. Y_{ig} is a continuous variable as it is derived from TPM values from Bayesian estimates of the transcript per million (TPM) value for all transcript species (isoforms) for a given gene. The distribution of Y_{ig} depends on hierarchically arranged parameters (Fig. 1), where each level in the hierarchy represents a different source of variation.

Y_{ig} is normally distributed with mean Z_{ig} and variance ϵ_{ig}. Z_{ig} is interpreted as allelic imbalance in expression, and the variation of Y_{ig} about Z_{ig} is considered to reflect a variation, or noise, which comes from both technical and biological source. The latter source induces some of the within-gene variation (across individuals) that is *not* explained by the experimentally controlled variables x_{ir} (age, sex, and type of cross, for which $r = 1, ..., 3$, respectively).

The remaining portion of within-gene, across individual, variation is framed in a normal linear model, where σ^2_{ig} also controls variation *not* explained by x_{ir}, whereas the regression parameters β_{gr} mediate the fixed effects of x_{ir} for $r = 1, ..., 3$.

The zeroth regression parameter β_{g0} may be interpreted as the overall propensity of gene g for allelic imbalance, when the effects of the three explanatory variables are averaged (hence β_{g0} is the intercept). Thus, for each gene g this propensity as well as the three effects yield four $\beta_{gr},\ r = 0, ..., 3$. Each of these varies across genes (i.e. transcriptome-wide), which is modeled by a mixture of a "narrow" and a c_r-times more varied normal distribution; and each is selected by a Bernoulli variable δ_{gr} with a gene-independent proportion p_r.

$\delta_{gr},\ r = 0, ..., 3$ are key variables, whose posterior probabilities are presented by Perez et al as informative summaries on each gene g (e.g. Fig. 1B of Perez et al). Most importantly, δ_{g0} indicates weather gene g is bi or monoallelically expressed (imprinted).

2 Comparison to the AGK models M1 and M2

A few overall features are shared between the AGK models (Fig. 2) and BRAIM. The most important are the independence of genes, the hierarchical structure separating distinct sources of biological and technical variation. But several important details differ (Table 1).

Firstly, the experimental design has two major differences, both being the consequence of human subjects: (i.) natural polymorphisms as genetic markers for maternal and paternal

[1] The authors themselves show that this is not so, but they keep the assumption for simplicity.

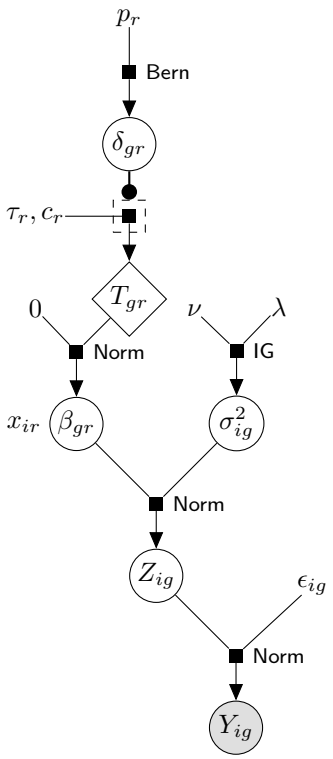

Figure 1: The BRAIM model

transcripts, and (ii.) uncontrolled explanatory variables (age of death, etc) exerting random effect.

Secondly, the observed variables Y_v are read counts, thus they are neither continuous nor do they contain all information that is relevant to a gene. This precludes direct application of BRAIM to our data. Moreover, Y_v is taken as certain observation ignoring much technical noise.

Thirdly, within-gene (across individual) variation is either entirely attributed to the measured explanatory variables x_{ir} (M2) or taken as fully independent of those (M1).

Fourthly, allelic imbalance is assumed to fall into a few discrete categories (two in the simplest, binary, case) instead of the fine-grained, continuous imbalance model of BRAIM. A related further limitation of M2 is that it cannot account for any across-genes variation in the effects of explanatory variables at a given imbalance category. Such variation has been found by Perez et al.

Fifthly, as opposed to the Bayesian BRAIM, the proposed inference based on AGK M1 and M2 is essentially frequentist (maximum likelihood estimation), which means that the parameters of interest such as π_k or β_{kr} are considered unknown but fixed quantities. A more practical consequence of frequentist inference is that it hinders inference based on more refined models (e.g. to allow effects to vary across genes), and may not work well even the relatively simple proposed AGK models.

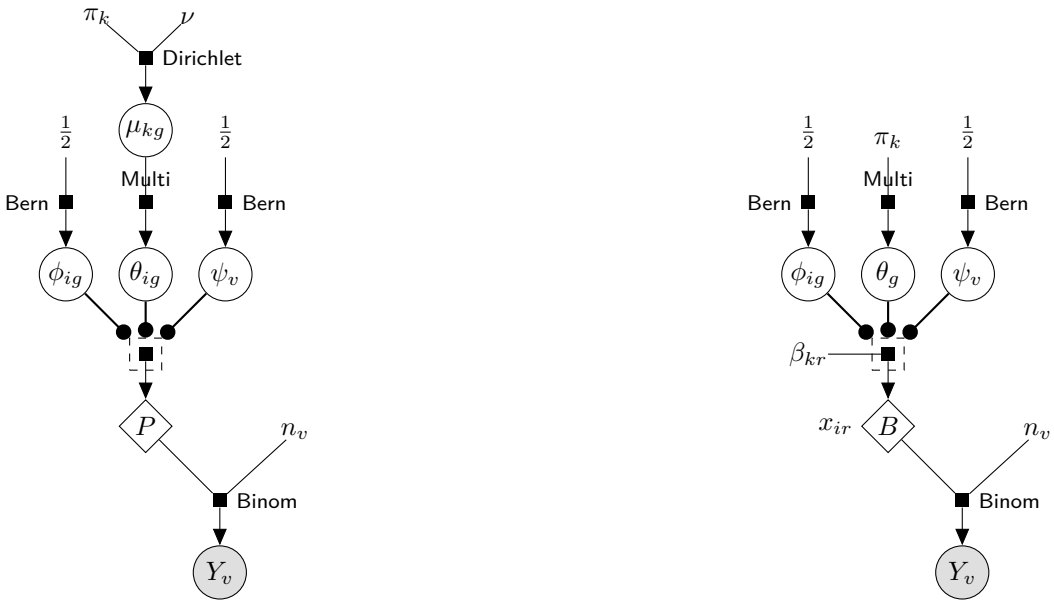

Figure 2: AGK models: M1 (left) and M2 (right)

	BRAIM	AGK M2
exp. design	mouse hybrid	humans
expl. var.	fixed effect	random effect
observed. var.	continuous TPM	read counts
observed. var. error	accounted for	ignored
across individ. var.	partially explained	fully explained
degree of imbalance	graded	discrete (categorical)
effects vary with genes	yes	not within a category
inference	Bayesian	frequentist

Table 1: Differences between BRAIM and AGK M2. The entries for AGK M1 would be identical to M2 except that the across individual variation is fully unexplained in M1.

3 Conclusion

BRAIM is more complex than any of the proposed AGK models but has greater descriptive power and allows more flexible inference, which may potentially open up new research directions. BRAIM was developed for a different experimental design so it is not directly applicable to our project. Nonetheless BRAIM could be adopted. The question is whether the amount of necessary work is worth the gain.

References

[1] Julio D Perez, Nimrod D Rubinstein, Daniel E Fernandez, Stephen W Santoro, Leigh A Needleman, Olivia Ho-Shing, John J Choi, Mariela Zirlinger, Shau-Kwaun Chen, Jun S Liu, and Catherine Dulac. Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. *eLife*, 4:e07860, jan 2015.