A General-er Coarse Data Model

Art Tay

Abstract

In the realm of high-throughput biological data, researchers are often faced with the challenge of integrating diverse measurement modalities, such as RNA-Seq and microarrays. These methods exhibit a trade-off between granularity and practicality: more granular techniques offer higher accuracy but come with increased costs and time requirements, while less granular methods are faster and more affordable but less precise. Despite these inherent differences, meta-analyses frequently necessitate the joint modeling of disparate data types to draw comprehensive conclusions. Traditional approaches, such as the classical coarse data model, have been applied in analogous contexts. However, these methods can falter when the fundamental assumption of sub-sample space inclusion is violated (Heitjan and Rubin 1991). For instance, while RNA-Seq provides precise count data, microarray experiments yield intensity values that may not correspond directly to any actual count, complicating integration efforts. To address this, we propose a novel Bayesian framework that incorporates a random coarsening function, enabling the joint modeling of heterogeneous data sources. Unlike previous Bayesian methods, our method explicitly models the functional relationship between RNA-Seq and microarray data (Ma et al. 2017). This allows researcher to directly applied any methods designed for singularly typed data. Ultimately, we believe that our method will be a more interpretable and general synthesis of different biological assays.

Preliminary Results

Assume that we have gene measurements on n samples in $X_{n\times p}$ where $p >> n$ with associated coarseness indicators $G_{n\times 1}$ and outcomes $Y_{n\times 1}$.

Data	Outcome	Gene 1	$\cdots$	Gene p	Coarseness
sample 1	0	count	$\cdots$	count	0
sample 2	1	continuous	$\cdots$	continuous	1
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
sample n	1	count	$\cdots$	count	0

Generally, we might be interested in modeling the data using a GLM

$$Y = g(X\beta) + \epsilon \quad \epsilon \sim f(\theta) \tag{1}$$

where f is some distribution with $\mathbb{E}[\epsilon] = 0$. Usually X is fixed making it easy to find the MLE for β; however, here X is random. Specifically, we will assume that:

$$X = \begin{cases} X & \text{if } G = 0 \\ h(X) & \text{if } G = 1 \end{cases} \tag{2}$$

where h is the unknown coarsening function. Now we can define the following hierarchical model:

$$
\begin{aligned}
L(\beta, h \mid Y, X, G) &\propto f(X, Y, G \mid \beta, h) \\
&\propto f(Y|X, G, \beta, h) \cdot p(X \mid G, h, \beta) \cdot f(\beta \mid h, G) \cdot f(h \mid G) \cdot f(G) \\
&\propto f(Y|X, G, \beta, h) \cdot p(X \mid G, h, \beta) \cdot f(\beta \mid h, G) \cdot f(h \mid G) \\
&\quad \text{in this context } G \text{ is known.} \\
&\propto f(Y|X, G, \beta, h) \cdot p(X \mid G, h) \cdot f(\beta) \cdot f(h \mid G) \\
&\quad \text{Assuming } h \text{ and } \beta \text{ are free of each other.}
\end{aligned} \tag{3}
$$

The likelihood $f(Y|X, G, \beta, h)$ is given by Eq. 1 and $p(X \mid G, h)$ is given by Eq. 2. Because $p >> n$ it is a good idea to specify a shrinkage prior on β such as $\mathcal{N}(0, \Sigma)$. In order to not have to specify the functional form of h, we will use a Gaussian process prior.

$$
f(h \mid G) \sim \begin{cases} X & \text{if } G = 0 \\ \mathcal{GP}(\mu_X, \Sigma_X) & \text{if } G = 1 \end{cases} \tag{4}
$$

References

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and coarse data," *The Annals of Statistics*, Institute of Mathematical Statistics, 19, 2244–2253.

Ma, T., Liang, F., Oesterreich, S., and Tseng, G. C. (2017), "A joint bayesian model for integrating microarray and RNA sequencing transcriptomic data," *Journal of Computational Biology*, 24, 647–662. https://doi.org/10.1089/cmb.2017.0056.