Using Approximate Message Passing for Low Rank Matrix Factorisation in Metagenomics

Jamie Dougherty*

Abstract

This work was done over the summer of 2018 for eight weeks under the supervision of Dr Sergio Bacallado at the Department of Pure Maths and Mathematical Statistics in the University of Cambridge. It was funded by the Summer Research in Mathematics Scheme at the department.

In this report I give a brief introduction the Metagenomic problem that was of interest. I provide an overview of the field of Approximate Message Passing and how it generalises to low-rank matrix factorisation. I show some of the work that I was able to accomplish when doing the project. I finish by providing some areas for future enquiry.

My hope is that with a mathematical background but who isn't familiar with any of the fields mentioned in this document will be able to understand the work that was done over this period.

1 Problem Outline

1.1 Metagenomics

Metagenomic experiments sequence the DNA of a mixture genomes, and frequently, the genomes present in each sample are not known. For example, a sample of bacterial DNA from the lungs of cystic fibrosis patients may contain many strains of the species Rothia mucilaginosa which have not been sequenced previously. Scientists are interested in (1): identifying these strains from metagenomic samples, and (2): understanding the evolution of the species in chronic infections. Learning more about these mixtures can open up new medical treatments such as the fecal microbiota transplant for Clostridium Difficile[3]. Conventional methods which involve intensive lab cultivation of samples don't scale well to these groups and so new methods are of interest.

1.2 Problem Specification

The problem of identifying new genomes from mixed samples is called de novo reconstruction[6]. This problem can be framed statistically as the reconstruction of a noisy low-rank matrix with structured factors. It has been shown that prior information on the factors of a matrix can significantly improve the error incurred by factorisation algorithms. However, it remains a challenge

*jamied157@gmail.com

to derive efficient algorithms which utilise such prior information, especially when the stochastic model for the factors is complex.[1]

I will now outline how the problem is constructed as a low rank matrix factorsation problem. We are given a collection of vectors $(\boldsymbol{z}_i)_{i=1}^n$, each representing a sample of genetic material extracted from the area of interest. Each vector then belongs to the space $[0,1]^{4m}$ where m is the number of genetic sites we have measure. The idea is that at each site we observe a certain number of nucleobases (A,T,C or G) and the elements correspond to the proportion of each nucleobase we observe. In practice there are a large number of sites that have the same base across all samples, these are removed because of their redundancy and so we will refer to the sites as *varying sites* to acknowledge this. The matrix $Z \in [0,1]^{n\times 4m}$ has rows set to be these vectors. The assumed factorisation is then

$$Z = UV + W, \quad U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{r\times 4m}$$

Where we have that W is a matrix of noise. U is a stochastic matrix (row-sums are equal to one) and

$$(V_{4i+j,k})_{j=0}^4 \in \{0,1\}^4 \quad \forall i,j$$

The idea here is that the matrix Z is actually generated by a small number r of populations of organisms (or taxa). The proportion of the j^{th} taxon in sample i is given by U_{ij}. The nucleobase of the i^{th} taxon in varying site j is given by V_{ij}. It's easy to see then why this factorisation is of interest - we learn more about the taxa that make up our mixture as well the proportions in this make-up.

1.3 Informative Priors

We can formulate our beliefs about the structure of the factors as prior distributions of the elements of the matrices, we have that

$$(U_{ij})_{i=1}^n \sim \text{Dirichlet}(1,1,\ldots,1)$$
$$(V_{4i+j,k})_{j=0}^4 \sim \text{Uniform on basis vectors}$$

To do a full Bayesian analysis we then need to describe how we observe these matrices by characterising the noise matrix W. For each sample i we observe each site d_i times and so we may have multinomial noise corrupting our observations. A normal assumption may suffice also.

From a mathematical point of view we may Here we have that $r << n, m$ and so we don't expect there to be many factorisations that satisfy this problem. The challenge rather is to find such a factorisation. The reason such a factorisation is of interest

2 Approximate Message Passing for Matrix Factorisation

Approximate Message Passing was introduced in [1] as a new iterative scheme similar to iterative thresholding algorithms to solve compressed sensing problems. The key difference between these

[1] may want to add more here later?

algorithms and the simple thresholding algorithms that had come before was the introduction of a correction term in these algorithms that improved performance significantly. As alluded to in the name, the inspiration for this change came from intuitions developed in statistical physics and graphical models literature from message passing schemes that try to perform inference on hidden variables under computational constraints. Further work then develop after this discovery, extending the algorithms to work in more general problems. Of particular interest to us are [7] and [5], the first allowed for noisy perturbations on the input and output of the desired vector while the second generalises it the the bilinear case. This allows for matrix factorisation problems of the kind specified above.

2.1 A brief introduction to AMP

Whilst a more thorough introduction exists in [1], a small introduction to AMP will be useful for the rest of this report. We are given the problem to find a sparse vector $\boldsymbol{x}$ when observing it passed through some linear system

$$\boldsymbol{y} = A\boldsymbol{x} \qquad A \in \mathbb{R}^{n \times N}, n < N$$

As we are transforming $\boldsymbol{x}$ into a lower dimension space we have a large number of solutions and so we need to incorporate the sparsity of $\boldsymbol{x}$ into our solution. A common scheme would be to use an iterative thresholding algorithm

$$\begin{aligned} \boldsymbol{x}^{t+1} &= \eta_t(A^T\boldsymbol{z}^t + \boldsymbol{x}^t) \\ \boldsymbol{z}^t &= \boldsymbol{y} - A\boldsymbol{x}^t \end{aligned}$$

Where η_t maps any element that exceed a certain amount to zero, hence the "thresholding" terminology. This is also how the scheme encourages sparsity. These updates are favoured in situations with large amounts of data as they are very computationally cheap. However the accuracy falls short of the best algorithms available. AMP alters this algorithm slightly

$$\begin{aligned} \boldsymbol{x}^{t+1} &= \eta_t(A^T\boldsymbol{z}^t + \boldsymbol{x}^t) \\ \boldsymbol{z}^t &= \boldsymbol{y} - A\boldsymbol{x}^t + \frac{1}{\delta} z^{t-1} \left\langle \eta_t'(A^T\boldsymbol{z}^{t-1} + \boldsymbol{x}^{t-1}) \right\rangle \end{aligned}$$

Here $\langle . \rangle$ denotes the empirical mean of a vector and $\eta'(s) = \frac{d}{ds}\eta(s)$. Note that this scheme still is computationally cheap however now the accuracy of the scheme reaches best in class. The term comes from the theory of belief-propagation and is dealt with more thoroughly in some of the papers referenced.

State Evolution

Another particularly attractive part of AMP is another set of iterative equations known as state evolution. These give an estimate for the error in each of the AMP estimates and allow the update equations to be tuned during the process. Proper treatments of this is found in [1] and [2].

2.2 Some examples

As a starting point for the project I reconstructed the results from the paper [4]. This provides a simple example to show how AMP works for matrix factorisation, I will provide a small outline of

the problem here and the reader may refer to the paper for more detail. We are given a matrix

$$A = \frac{\lambda}{n} \boldsymbol{x}_0 \boldsymbol{x}_0^T$$

For some vector $\boldsymbol{x}_0$ generated from a simple distribution. However, the results in the paper are valid for any $\boldsymbol{x}_0$ as long as the asymptotic distribution of the $\boldsymbol{x}_0$ is known. We then define the recursion

$$\boldsymbol{x}^{t+1} = A f_t(\boldsymbol{x}^t) - b_t f_{t-1}(\boldsymbol{x}^{t-1})$$

Where the $\boldsymbol{x}^t$ is an estimate for $\boldsymbol{x}_0$. Whilst the definition of f_t is somewhat technical it is sufficient to think of it as a denoising function that tries to push the estimate towards areas of higher probability in $\boldsymbol{x}_0$'s distribution.

TBC

References

[1] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[2] A. Javanmard and A. Montanari. State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling. *arXiv e-prints*, November 2012.

[3] Sonny TM Lee, Stacy A Kahn, Tom O Delmont, Nathaniel A Hubert, Hilary G Morrison, Dionysios A Antonopoulos, David T Rubin, and A. Murat Eren. High-resolution tracking of microbial colonization in fecal microbiota transplantation experiments via metagenome-assembled genomes. *bioRxiv*, 2016.

[4] A. Montanari and R. Venkataramanan. Estimation of Low-Rank Matrices via Approximate Message Passing. *arXiv e-prints*, November 2017.

[5] Jason T. Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing. *CoRR*, abs/1310.2632, 2013.

[6] Christopher Quince, Tom O Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E Darling, Gavin Collins, and A Murat Eren. Desman: a new tool for de novo extraction of strains from metagenomes. *Genome biology*, 18(1):181, 2017.

[7] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. *CoRR*, abs/1010.5141, 2010.