Module 3: Copy number variation

5th June 2019

Overview

- Copy number variants (CNVs) are another common genomic aberration in cancer.
- Like SNVs they can act as driver mutations and they can be markers of clonal populations.
- The main difference between CNVs and SNVs is spatial correlation.

Data representation

- What data should we use to model CNVs?
- We assume HTS data, either bulk or single cell.
- We can summarise aligned reads in multiple ways for CNV analysis.
  - Allele counts
  - Binned read counts
  - Haplotype blocks

Allelic counts

- Like SNVs we can consider allelic counts.
- Most positions won't have a SNV or SNP so we mainly have information about total depth.
- If we restrict to heterozygous SNPs we can have read counts for the reference *a* and alternate allele *b*.
- This representation is good for allele specific copy number inference.
- It is not a useful for low coverage data like single WGS.

Binned read counts

- Instead of looking point wise at genomic positions, we divide the genome into bins.
  - Typical bin lengths are of the order 10^3bp – 10^5bp.
  - Ideally bin length should be smaller than the typical CNV size.
- We can then count the number of reads in the bin.
- Assuming read depth is proportional to copy number we can then model total copy number.
- This strategy is useful in low coverage data as we pool data across much larger regions.
- It is also computationally cheaper as we have 10^4 – 10^6 data points.

Haplotype blocks

- A haplotype block is a region of the genome where we can phase heterozygous SNPs.
  - That is we can determine the parental sequence of SNPs on each chromosome.
- We can summarise the read data from these blocks in terms of three values.
  - Total number of reads.
  - Number of reads supporting haplotype 1 at het SNPs.
  - Number of reads supporting haplotype 2 at het SNPs.
- For bulk data this representation is typically an improvement over the allelic counts.
- It is less useful for low coverage data unless we can get reasonably long haplotype blocks.

Modelling spatial correlation

- The key point for CNVs is we want to model spatial correlation.
  - Two adjacent points in the chromosome are more likely to have the same copy number.
- There are two major strategies employed:
  1. Segment the data in advance and then fit the copy number.
  2. Jointly segment the data and fit copy number.
- Strategy 1 is computationally efficient. However, the segmentation is fixed and thus cannot make use of information such as tumour content.
- We will explore strategy 2 here.

Question
Does anyone know of tools which use each strategy?

Probabilistic modelling of spatial correlation

- Hidden Markov models (HMMs) are the most popular approach for modelling spatial correlation in computational cancer.
- Other approaches such as Kalman filtering and general state space models have not been widely used.
- HMMs assume a data point only depends on the point immediately preceding it.
  - Important for developing efficient inference algorithms.
- Downside of HMMs is limited ability to model state duration.

HMMs

- Latent variable model, where the latent (or hidden) variables have a dependency structure.
- Latent variables take values in some discrete state space $\mathcal{X}$.
- Dynamic mixture models, where mixture weights depend on previous data point.

HMMs

- π - The initial state vector. This is a vector of length $|\mathcal{X}|$ where π_i is the probability the first hidden variable in the sequence takes value i.
- A - The transition matrix. This is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix where A_{ij} is the probability that the current hidden value takes on state j given the previous one was in state i.
- F - The emission distribution. This is the distribution for the observed data which depends on the associated hidden state. The hidden state is like the cluster indicator in mixture models, selecting which parameter is used for F to generate the data point.
- Here we assume that $K = |\mathcal{X}|$ is known and fixed. This assumption can be relaxed using non-parametric Bayesian priors.
  - Computationally demanding.

Bayesian HMMs

$$
\begin{aligned}
\boldsymbol{\pi}|\boldsymbol{\kappa} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\kappa}) \\
A_{k\cdot}|\boldsymbol{\gamma} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\gamma}) \\
z_1|\boldsymbol{\pi} &\sim \text{Categorical}(\cdot|\boldsymbol{\pi}) \\
z_n|z_{n-1}, A &\sim \text{Categorical}(\cdot|A_{z_{n-1}\cdot}) \\
\theta_k &\sim G \\
x_n|z_n, \boldsymbol{\theta} &\sim F(\cdot|\theta_{z_n})
\end{aligned}
$$

HMM inference

- We can efficiently compute the conditional distribution of the hidden states.
  - Done using the forward-backward (FB) algorithm
- FB can be used for both MAP (via EM) and MCMC estimation.
- Replacing summation with maximisation in FB leads to Viterbi algorithm.
  - Compute the most probable sequence of states.
- We discuss using EM to perform MAP.
  - Common approach in the field due to computational complexity of MCMC.
  - We use EM to find the MAP values of π and A then Viterbi to find the hidden states.

Forward-Backward recursion

$$\alpha(z_n = k) = p(x_n|z_n = k) \sum_{l=1}^{K} \alpha(z_{n-1} = l) A_{lk}$$

$$\beta(z_n = k) = \sum_{l=1}^{K} \beta(z_{n+1} = l) p(x_{n+1}|z_{n+1} = l) A_{lk}$$

E-step

$$
\begin{aligned}
p(\mathbf{x}) &= \sum_{k=1}^{K} \alpha(Z_n = K) \\
\mathbb{E}_{\mathbf{z}}[\mathbb{I}(z_n = k)] &= \frac{\alpha(z_n = k)\beta(z_n = k)}{p(\mathbf{x})} \\
\mathbb{E}_{\mathbf{z}}[\mathbb{I}(z_{n-1} = k, z_n = l)] &= \frac{\alpha(z_n = k)p(x_{n-1}|z_{n-1} = k)p(x_n|z_n = l)\beta(z_n = l)}{p(\mathbf{x})}
\end{aligned}
$$

M-step

$$\hat{\pi}_k \propto \kappa_k + \mathbb{E}_{\mathbf{z}}[\mathbb{I}(z_n = k)]$$

$$\hat{A}_{kl} \propto \gamma_{kl} + \mathbb{E}_{\mathbf{z}}[\mathbb{I}(z_{n-1} = k, z_n = l)]$$

$$\frac{\partial}{\partial \theta_k} \log p(\mathbf{x}, \boldsymbol{\pi}, A, \boldsymbol{\theta}) = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}}[\mathbb{I}(z_n = k)] \frac{\partial}{\partial \theta_k} \log f(x_n|\theta_k) + \frac{\partial}{\partial \theta_k} \log p(\theta_k)$$

CNV modelling

Homogeneous sample model

- We consider the problem of inferring copy number profiles from a sample that has:
  - No clonal population structure.
  - No normal contamination.
- Data is binned total read counts.
- We want to infer total copy number.
- While a bit uninteresting for bulk, it could be useful for single cell.

Notation

- m - chromosome index
- M - number of chromosomes
- n - bin index
- N_m - number of bins for chromosome m
- x_n^m - number of reads in the n^{th} bin on chromosome m
- z_n^m - hidden state for the n^{th} bin on chromosome m

Model

$$
\begin{aligned}
\boldsymbol{\pi}|\boldsymbol{\kappa} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\kappa}) \\
A_{k\cdot}|\boldsymbol{\gamma} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\gamma}) \\
z_1^m|\boldsymbol{\pi} &\sim \text{Categorical}(\cdot|\boldsymbol{\pi}) \\
z_n^m|z_{n-1}^m, A &\sim \text{Categorical}(\cdot|A_{z_{n-1}^m\cdot}) \\
r &\sim \text{Gamma}(\cdot|a,b) \\
\theta_c|r &= rc \\
x_n^m|z_n^m, \boldsymbol{\theta} &\sim \text{Poisson}(\cdot|\theta_{z_n^m})
\end{aligned}
$$

- r - haploid read depth
- θ_c - deterministic function of copy number and read depth

GC content correction

- Simple model assumes read depth is a linear function of copy number and haploid coverage.
- In practice the GC content of a bin has a substantial impact on read coverage.
- One strategy is to normalise for this a pre-processing.
  - Data is no longer integer and Poisson cannot be used.
- Alternative is to introduce GC covariate g_n^m and model its effect

$$\begin{aligned} h_n^m &\sim H(\cdot|g_n^m) \\ x_n^m|h_n^m, z_n^m, \boldsymbol{\theta} &\sim \text{Poisson}(\cdot|h_n^m \theta_{z_n^m}) \end{aligned}$$

Overdispersion

- Like the SNV case, data can have more variance then our assumed distribution.
- We can follow the same strategy as before and replace the Poisson with a Negative-Binomial with the same mean.
  - We now have an additional variance parameter which could be global or state specific.
- Unlike SNV data, overdispersion will appear in WGS data. This is because binning leads to much higher read depth so the effect become apparent.

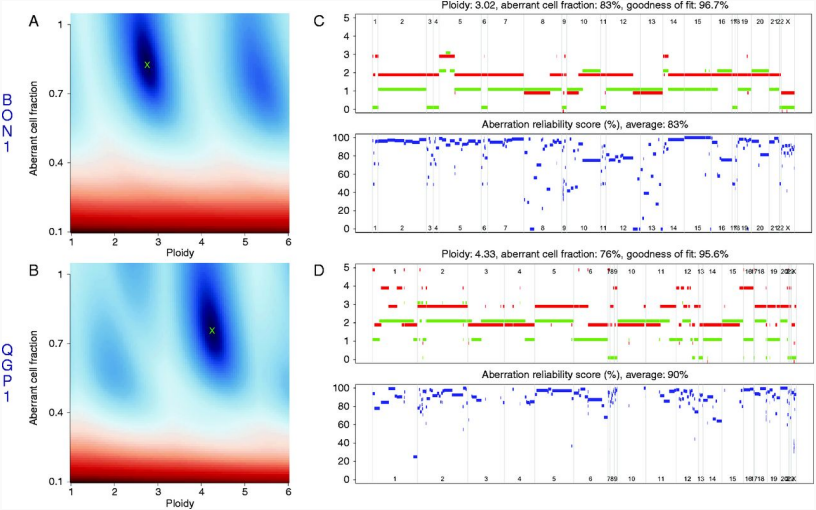
Normal contamination

$$
\begin{aligned}
\boldsymbol{\pi}|\boldsymbol{\kappa} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\kappa}) \\
A_{k\cdot}|\boldsymbol{\gamma} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\gamma}) \\
z_1^m|\boldsymbol{\pi} &\sim \text{Categorical}(\cdot|\boldsymbol{\pi}) \\
z_n^m|z_{n-1}^m, A &\sim \text{Categorical}(\cdot|A_{z_{n-1}^m\cdot}) \\
r &\sim \text{Gamma}(\cdot|a,b) \\
t &\sim \text{Uniform}(\cdot|[0,1]) \\
\theta_c|r,t &= r(2(1-t)+ct) \\
x_n^m|z_n^m,\boldsymbol{\theta} &\sim \text{Poisson}(\cdot|\theta_{z_n^m})
\end{aligned}
$$

Ploidy

- Model is unidentifiable because of $\theta_c = rc$.
  - Doubling c and halving r leads to solution with same likelihood
- Ploidy is loosely used to refer to the average copy number of a sample (cell).
- We can alter the ploidy and change r to come up with multiple equally likely solutions.
- Problem for all CNV software, bulk or single cell.
- No automated solution - manual curation required

Ploidy example (ASCAT)



Vandamme et al. Journal of Molecular Endocrinology 2015

Allele specific copy number

- If we want to infer major and minor copy number we need to incorporate het SNP information.
- Using haplotype block data is better than allele specific counts.
  - Pooling strength across blocks allows for better estimates of allele frequencies.
- x_n^a and x_n^b be the number of reads supporting haplotype a and b in the block.
- x_n^t is the total number of reads in the block.
- x_n^l is the length of the block
- $z_n \in \{(c_a, c_b) : c_a, c_b \leq C\}$

Allele specific copy number

$$
\begin{aligned}
\boldsymbol{\pi}|\boldsymbol{\kappa} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\kappa}) \\
A_{k\cdot}|\boldsymbol{\gamma} &\sim \text{Dirichlet}(\cdot|\boldsymbol{\gamma}) \\
z_1|\boldsymbol{\pi} &\sim \text{Categorical}(\cdot|\boldsymbol{\pi}) \\
z_n|z_{n-1}, A &\sim \text{Categorical}(\cdot|A_{z_{n-1}\cdot}) \\
r &\sim \text{Gamma}(\cdot|a,b) \\
\theta_c|r, z_n = (c_a, c_b), x_n^l &= r(c_a + c_b)x_n^l \\
x_n^t|z_n^m, \boldsymbol{\theta} &\sim \text{Poisson}(\cdot|\theta_{z_n}) \\
x_n^b|x_n^a, z_n = (c_a, c_b) &\sim \text{Binomial}(\cdot|x_n^a + x_n^d, \frac{c_b}{c_a + c_b})
\end{aligned}
$$