Extracting structure from contaminated symbolic data

Antony Pearson

Applied Mathematics, University of Colorado Boulder
IQ Biology Program, BioFrontiers Institute

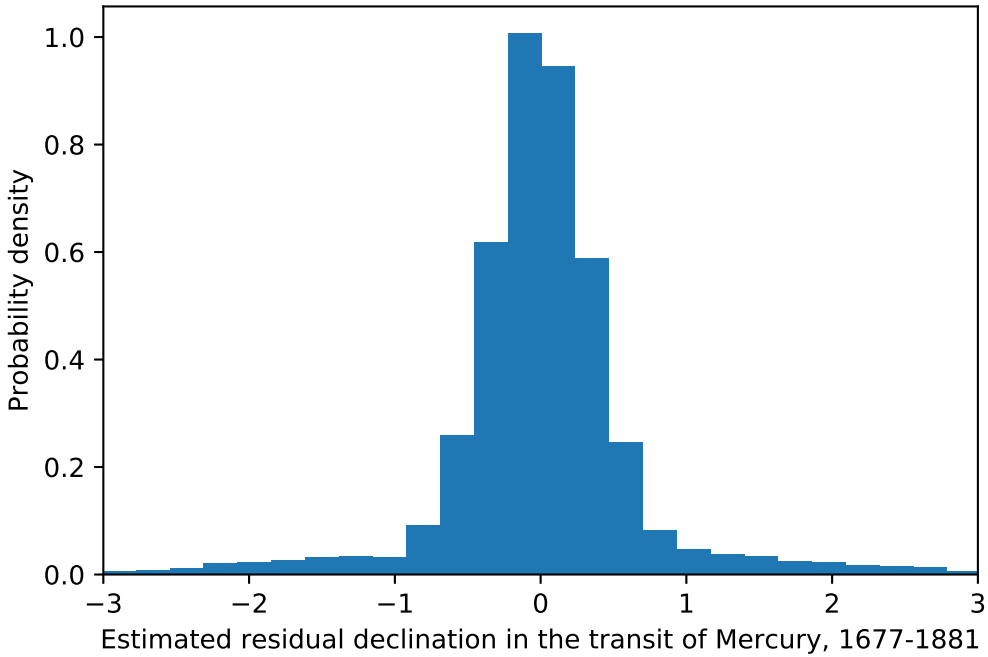
10-th ACM BCB Conference
Niagara Falls, NY

A Jupyter notebook and associated data are available at

`https://github.com/antonypearson/ACM-BCB-2019-Contamination`

It may be necessary to unzip some files.

Outliers and contamination



Outliers and contamination

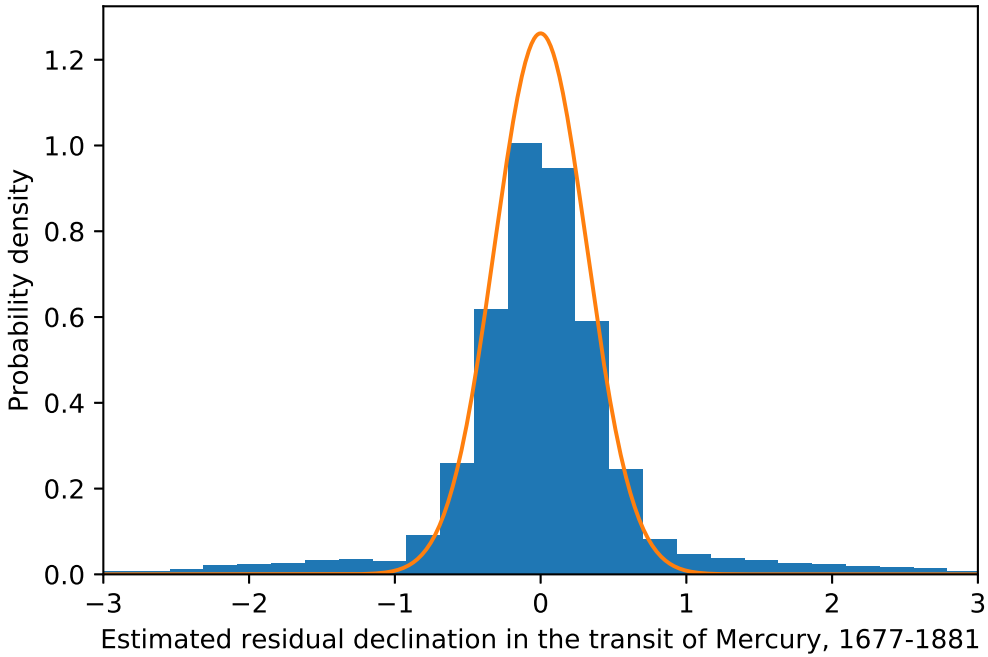
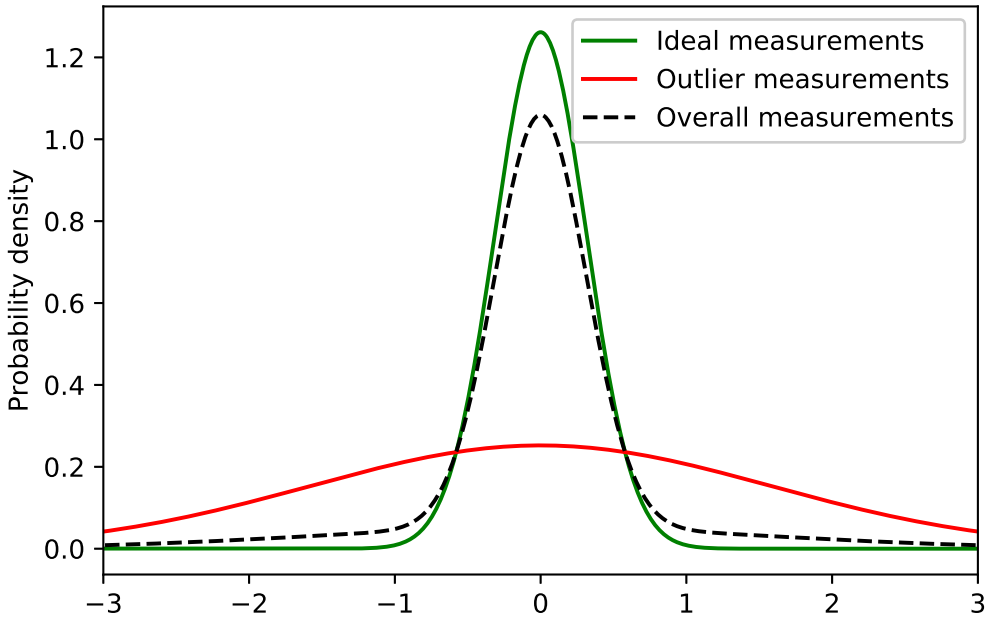


Figure: Heavy tails! But why?

Outliers and contamination



A common contamination model

When continuous data don't conform exactly to an idealised model (e.g. Gaussian), contamination models can be used to describe deviation from the theoretical model. Contamination models are mixture models:

A common contamination model

When continuous data don't conform exactly to an idealised model (e.g. Gaussian), contamination models can be used to describe deviation from the theoretical model. Contamination models are mixture models:

$$P = (1 - \lambda) \cdot R + \lambda \cdot P^*.$$

A common contamination model

When continuous data don't conform exactly to an idealised model (e.g. Gaussian), contamination models can be used to describe deviation from the theoretical model. Contamination models are mixture models:

$$P = (1 - \lambda) \cdot R + \lambda \cdot P^*.$$

"Robust statistics" can minimize the effect of the contaminating data on parameter estimation, and contamination models can be estimated under strong assumptions on R, e.g. P^* and R are both Gaussian with identical mean.

Coding section 1

Symbolic data in biology

Contemporary biological research produces data of a different kind. In "omics data" random variables do not necessarily take values in a real number space, and a notion of normally-distributed error often does not make sense. For example:

Symbolic data in biology

Contemporary biological research produces data of a different kind. In "omics data" random variables do not necessarily take values in a real number space, and a notion of normally-distributed error often does not make sense. For example:

- DNA and RNA sequencing reads

Symbolic data in biology

Contemporary biological research produces data of a different kind. In "omics data" random variables do not necessarily take values in a real number space, and a notion of normally-distributed error often does not make sense. For example:

- DNA and RNA sequencing reads
- DNA methylation sequencing

Symbolic data in biology

Contemporary biological research produces data of a different kind. In "omics data" random variables do not necessarily take values in a real number space, and a notion of normally-distributed error often does not make sense. For example:

- DNA and RNA sequencing reads
- DNA methylation sequencing
- Protein sequences

Symbolic data in biology

Contemporary biological research produces data of a different kind. In "omics data" random variables do not necessarily take values in a real number space, and a notion of normally-distributed error often does not make sense. For example:

- DNA and RNA sequencing reads
- DNA methylation sequencing
- Protein sequences
- Offspring phenotype counts

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions

Common assumptions on symbolic data

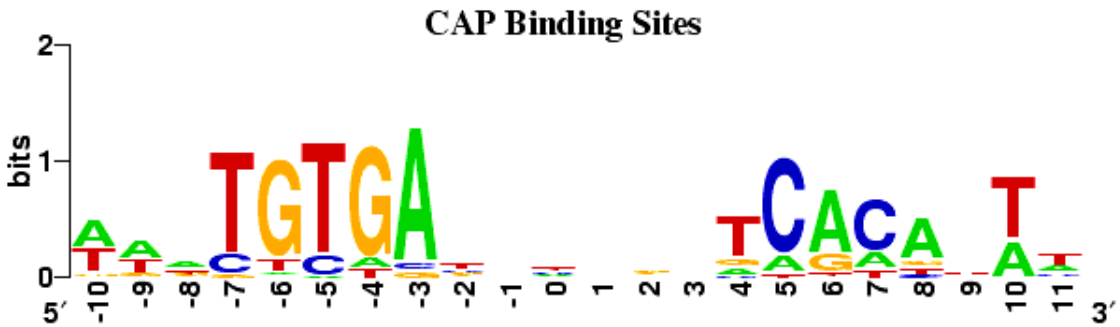
While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)



Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation
  - The number of sequencing reads inside a genomic window of fixed size

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation
  - The number of sequencing reads inside a genomic window of fixed size
- "i.i.d.-ness"

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation
  - The number of sequencing reads inside a genomic window of fixed size
- "i.i.d.-ness"
  - Random DNA sequences and E-values

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation
  - The number of sequencing reads inside a genomic window of fixed size
- "i.i.d.-ness"
  - Random DNA sequences and E-values
  - Substitutions in non-functional regions of DNA

Common assumptions on symbolic data

While assumptions like Gaussian distribution are natural with many continuous datasets, some things commonly assumed about symbolic data are:

- Independence
  - The number of crossovers in non-overlapping genomic regions
  - Transcription factor binding sites (logo plots)
- Poisson law
  - The number of mutations per generation
  - The number of sequencing reads inside a genomic window of fixed size
- "i.i.d.-ness"
  - Random DNA sequences and E-values
  - Substitutions in non-functional regions of DNA
- Exchangeability

Understanding models of contamination

- Literally, "contamination" (of a biological sample)

Understanding models of contamination

- Literally, “contamination” (of a biological sample)
- Human error

Understanding models of contamination

- Literally, "contamination" (of a biological sample)
- Human error
- Systematic error (e.g. base substitutions in sequencing)

Understanding models of contamination

- Literally, "contamination" (of a biological sample)
- Human error
- Systematic error (e.g. base substitutions in sequencing)
- Hidden biological mechanism which slightly breaks assumptions

Contamination in symbolic data

- *Sample space* Ω: finite set — e.g. the methylation configurations of three contiguous "CpGs" in DNA ($\{(0,0,0),(0,0,1),\ldots\}$

Contamination in symbolic data

- *Sample space* Ω: finite set — e.g. the methylation configurations of three contiguous "CpGs" in DNA ($\{(0,0,0),(0,0,1),\ldots\}$
- $\mathcal{P}$, the set of all probability models over Ω — can be thought of as a non-negative vector which sums to 1

Contamination in symbolic data

- *Sample space* Ω: finite set — e.g. the methylation configurations of three contiguous "CpGs" in DNA ($\{(0,0,0),(0,0,1),\ldots\}$
- $\mathcal{P}$, the set of all probability models over Ω — can be thought of as a non-negative vector which sums to 1
- A specially-structured class of probability models $\mathcal{Q}$, having desirable properties — e.g. Poisson law or independence

Given any probabilistic model P in $\mathcal{P}$, *how much data from P can be attributed to a well-structured model in $\mathcal{Q}$?*

What form does contamination take?

$$P = \lambda \cdot Q + (1 - \lambda) \cdot R$$

Unlike continuous data, where Gaussian contamination is often phenomonologically well-justified, if Ω is e.g. the set of DNA k-mers, there is no clear structure that R should follow.

Requiring R to be the uniform distribution over all k-mers, for instance, would imply strong beliefs about the mechanism that causes errors in sequencing.

What form does contamination take?

$$P = \lambda \cdot Q + (1 - \lambda) \cdot R$$

Unlike continuous data, where Gaussian contamination is often phenomonologically well-justified, if Ω is e.g. the set of DNA k-mers, there is no clear structure that R should follow.

Requiring R to be the uniform distribution over all k-mers, for instance, would imply strong beliefs about the mechanism that causes errors in sequencing.

This motivates us to deconstrain the structure of contamination and instead focus on the well-structured model hidden inside the model.

Focusing on purity, not contamination

For a class of well-structured models $\mathcal{Q}$, we aim to know what proportion data (from the model P) can be attributed to a model Q in $\mathcal{Q}$.

That is, we want a representation of the form:

$$P = \lambda \cdot Q + (1 - \lambda) \cdot R,$$

where λ is as large as possible, and R is an arbitrary probability model (usually not in $\mathcal{Q}$).

Focusing on purity, not contamination

For a class of well-structured models $\mathcal{Q}$, we aim to know what proportion data (from the model P) can be attributed to a model Q in $\mathcal{Q}$.

That is, we want a representation of the form:

$$P = \lambda \cdot Q + (1 - \lambda) \cdot R,$$

where λ is as large as possible, and R is an arbitrary probability model (usually not in $\mathcal{Q}$).

For instance, if Q's sought structure is Poisson, on average $(100 \cdot \lambda)\%$ of samples from P will appear to follow a Poisson distribution.

The latent weight of $\mathcal{Q}$ in P

For any well-structured class $\mathcal{Q}$ and any model P, define

$$\lambda_{\mathcal{Q}}(P) := \begin{cases} \text{the largest } \lambda \text{ such that } P \geq \lambda \cdot Q \\ \text{for some well-structured model } Q \end{cases}$$

We call this the *latent weight of $\mathcal{Q}$ in P*

Then:

The latent weight of $\mathcal{Q}$ in P

For any well-structured class $\mathcal{Q}$ and any model P, define

$$\lambda_{\mathcal{Q}}(P) := \begin{cases} \text{the largest } \lambda \text{ such that } P \geq \lambda \cdot Q \\ \text{for some well-structured model } Q \end{cases}$$

We call this the *latent weight of $\mathcal{Q}$ in P*

Then:

- $0 \leq \lambda_{\mathcal{Q}}(P) \leq 1$

The latent weight of $\mathcal{Q}$ in P

For any well-structured class $\mathcal{Q}$ and any model P, define

$$\lambda_{\mathcal{Q}}(P) := \begin{cases} \text{the largest } \lambda \text{ such that } P \geq \lambda \cdot Q \\ \text{for some well-structured model } Q \end{cases}$$

We call this the *latent weight of $\mathcal{Q}$ in P*

Then:

- $0 \leq \lambda_{\mathcal{Q}}(P) \leq 1$
- If $\mathcal{Q}$ is just one specific model Q, $\lambda_{\mathcal{Q}}(P)$ is the minimum of the ratio $P(\omega)/Q(\omega)$ over all possible outcomes ω

The latent weight of $\mathcal{Q}$ in P

For any well-structured class $\mathcal{Q}$ and any model P, define

$$\lambda_{\mathcal{Q}}(P) := \begin{cases} \text{the largest } \lambda \text{ such that } P \geq \lambda \cdot Q \\ \text{for some well-structured model } Q \end{cases}$$

We call this the *latent weight of $\mathcal{Q}$ in P*

Then:

- $0 \leq \lambda_{\mathcal{Q}}(P) \leq 1$
- If $\mathcal{Q}$ is just one specific model Q, $\lambda_{\mathcal{Q}}(P)$ is the minimum of the ratio $P(\omega)/Q(\omega)$ over all possible outcomes ω
- In general, $\lambda_{\mathcal{Q}}(P)$ is the largest that $\min_{\omega} \frac{P(\omega)}{Q(\omega)}$ can be, with Q a member of the structured class $\mathcal{Q}$

The latent weight of $\mathcal{Q}$ in P

Typically:

$$P = \lambda_{\mathcal{Q}}(P) \cdot Q + (1 - \lambda_{\mathcal{Q}}(P)) \cdot R,$$

for some Q in $\mathcal{Q}$ and some usually unstructured probabilistic model R

The latent weight of $\mathcal{Q}$ in P

Typically:

$$P = \lambda_{\mathcal{Q}}(P) \cdot Q + (1 - \lambda_{\mathcal{Q}}(P)) \cdot R,$$

for some Q in $\mathcal{Q}$ and some usually unstructured probabilistic model R

That is, we expect $100\lambda_{\mathcal{Q}}(P)\%$ of samples from P will appear to originate from a model in $\mathcal{Q}$!

A remarkable example

Suppose $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ i.e. $\Omega = \{0,1\}^2$

Consider the joint p.m.f. of binary random variables X, Y:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

A remarkable example

Suppose $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ i.e. $\Omega = \{0,1\}^2$

Consider the joint p.m.f. of binary random variables X, Y:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

(For example, 0 could represent a purine and 1 a pyrimidine, and X and Y are adjacent nucleotides.)

A remarkable example

Suppose $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ i.e. $\Omega = \{0,1\}^2$

Consider the joint p.m.f. of binary random variables X, Y:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

(For example, 0 could represent a purine and 1 a pyrimidine, and X and Y are adjacent nucleotides.)

Marginally, $X \sim$Bernoulli(0.8) and $Y \sim$Bernoulli(0.6), however:

$$P \neq \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.2 \times 0.4 & 0.8 \times 0.4 \\ 0.2 \times 0.6 & 0.8 \times 0.6 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

i.e. X and Y are not independent

A remarkable example

Suppose $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ i.e. $\Omega = \{0,1\}^2$

Consider the joint p.m.f. of binary random variables X, Y:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

(For example, 0 could represent a purine and 1 a pyrimidine, and X and Y are adjacent nucleotides.)

Marginally, $X \sim$Bernoulli(0.8) and $Y \sim$Bernoulli(0.6), however:

$$P \neq \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.2 \times 0.4 & 0.8 \times 0.4 \\ 0.2 \times 0.6 & 0.8 \times 0.6 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

i.e. X and Y are not independent

A remarkable example

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

Do data from P appear independent, however? Let's simulate data from P and test the hypothesis that X and Y are independent.

Coding section 2

A remarkable example (cont)

Recall:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

What is the latent weight of the independent models in P?

A remarkable example (cont)

Recall:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

What is the latent weight of the independent models in P?

We need to compute:

$$\lambda_{\mathcal{Q}}(P) = \max_{Q} \min_{\omega \in \Omega} \frac{P(\omega)}{Q(\omega)}$$

where the max is taken over all possible independent models Q with sample space $\Omega = \{0,1\}^2$

Coding section 3

A remarkable example (cont)

Recall:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

What is the latent weight of the independent models in P?

A remarkable example (cont)

Recall:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

What is the latent weight of the independent models in P?

In fact, $\lambda_{\mathcal{Q}}(P) = 96\%$, where Q is the independent model that renders $X \sim$Bernoulli$(5/6)$ and $Y \sim$Bernoulli$(5/8)$

A remarkable example (cont)

Recall:

$$P = \begin{array}{c} \begin{array}{cc} X=0 & X=1 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 \\ 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} Y=0 \\ Y=1 \end{array}$$

What is the latent weight of the independent models in P?

In fact, $\lambda_{\mathcal{Q}}(P) = 96\%$, where Q is the independent model that renders $X \sim$Bernoulli$(5/6)$ and $Y \sim$Bernoulli$(5/8)$

(The independent model formed from the marginals of X and Y only has weight $\approx 83\%$ as a component of P, which seems relatively small.)

Connections to hypothesis testing

The traditional approach to understanding the structure of a probabilistic model is to collect a sample and test the hypotheses:

$H_0 : P$ belongs to the structured class $\mathcal{Q}$ of models

$H_1 : P$ does not have $\mathcal{Q}$'s special structure

Connections to hypothesis testing

The traditional approach to understanding the structure of a probabilistic model is to collect a sample and test the hypotheses:

$$H_0 : P \text{ belongs to the structured class } \mathcal{Q} \text{ of models}$$
$$H_1 : P \text{ does not have } \mathcal{Q}\text{'s special structure}$$

p-values are sample dependent; in particular, are random quantities

Connections to hypothesis testing

The traditional approach to understanding the structure of a probabilistic model is to collect a sample and test the hypotheses:

$$H_0 : P \text{ belongs to the structured class } \mathcal{Q} \text{ of models}$$
$$H_1 : P \text{ does not have } \mathcal{Q}\text{'s special structure}$$

p-values are sample dependent; in particular, are random quantities

In contrast, the latent weight of $\mathcal{Q}$ in P, i.e. $\lambda_{\mathcal{Q}}(P)$, is an intrinsic property of P regardless whether H_0 or H_1 is true, and can be understood outside the context of a sample from P

Connections to hypothesis testing

- When H_1 is true but $\lambda_{\mathcal{Q}}(P) \approx 1$, it may require a very large sample to reject H_0, leading one to likely accept a false null hypothesis

Connections to hypothesis testing

- When H_1 is true but $\lambda_{\mathcal{Q}}(P) \approx 1$, it may require a very large sample to reject H_0, leading one to likely accept a false null hypothesis

- Conversely, knowing $\lambda_{\mathcal{Q}}(P) \approx 1$ may save us from using a needlessly complex model

  For instance, perhaps one collects a very large sample from P and blindly rejects the hypothesis of independence based on the p-value, when for their application it is acceptable to model the random variables more simply as they are independent 96% of the time.

Difficulties with the numerical approach

In the previous example we created a grid with 1% accuracy in each of two variable over the independent binary models, a relatively small space

Difficulties with the numerical approach

In the previous example we created a grid with 1% accuracy in each of two variable over the independent binary models, a relatively small space

If we try to approximate $\lambda_{\mathcal{Q}}(P)$ when $\mathcal{Q}$ is the set of independent 4-dimensional binary model (e.g. a 4-mer being independent in the sequence of its purines and pyrimidines), we would have to compute

$$\min_{\omega \in \{0,1\}^4} \frac{P(\omega)}{Q(\omega)}$$

for 100 million distributions Q. This is prohibitive!

More explicit descriptions of latent weights

To avoid computational issues it may be necessary to derive a more explicit form of $\lambda_{\mathcal{Q}}(P)$ that takes advantage of $\mathcal{Q}$'s structure

More explicit descriptions of latent weights

To avoid computational issues it may be necessary to derive a more explicit form of $\lambda_{\mathcal{Q}}(P)$ that takes advantage of $\mathcal{Q}$'s structure

To this end the following identity may be useful:

$$\begin{aligned}\lambda_{\mathcal{Q}}(P) &= \max_{Q \in \mathcal{Q}} \min_{\omega \in \Omega} \frac{P(\omega)}{Q(\omega)} \\ &= \max_{\omega \in \Omega} \left\{ \max_{Q \in \mathcal{Q}_\omega} \frac{P(\omega)}{Q(\omega)} \right\}\end{aligned}$$

More explicit descriptions of latent weights

To avoid computational issues it may be necessary to derive a more explicit form of $\lambda_{\mathcal{Q}}(P)$ that takes advantage of $\mathcal{Q}$'s structure

To this end the following identity may be useful:

$$\begin{aligned}\lambda_{\mathcal{Q}}(P) &= \max_{Q \in \mathcal{Q}} \min_{\omega \in \Omega} \frac{P(\omega)}{Q(\omega)} \\ &= \max_{\omega \in \Omega} \left\{ \max_{Q \in \mathcal{Q}_\omega} \frac{P(\omega)}{Q(\omega)} \right\}\end{aligned}$$

Here, $\mathcal{Q}_\omega$ denotes all the models Q in $\mathcal{Q}$ which satisfy:

$$\frac{P(\omega)}{Q(\omega)} \leq \frac{P(\omega')}{(Q\omega')}$$

for every other outcome ω' in the sample space Ω

Explicit form of independent weights

To fix ideas, take $\mathcal{Q}$ to represent distribution of independent d-dimensional binary variables. Then

$$\lambda_{\mathcal{Q}}(P) = \max_{\omega \in \{0,1\}^d} \max_{Q \in \mathcal{Q}_\omega \leftrightarrow (q_1,\ldots,q_d)} \frac{P(\omega)}{\prod_{i=1}^{d} q_i^{\omega_i}(1-q_i)^{1-\omega_i}}$$

Explicit form of independent weights

To fix ideas, take $\mathcal{Q}$ to represent distribution of independent d-dimensional binary variables. Then

$$\lambda_{\mathcal{Q}}(P) = \max_{\omega \in \{0,1\}^d} \max_{Q \in \mathcal{Q}_\omega \leftrightarrow (q_1,\ldots,q_d)} \frac{P(\omega)}{\prod_{i=1}^{d} q_i^{\omega_i}(1-q_i)^{1-\omega_i}}$$

For instance, when $d = 2$ and $\omega = (0,0)$, we need to determine non-negative q_1, q_2 such that:

- $q_1 + q_2 = 1$
- $\frac{p_{(0,0)}}{(1-q_1)(1-q_2)} \leq \frac{p_{(0,1)}}{(1-q_1)q_2}$
- $\frac{p_{(0,0)}}{(1-q_1)(1-q_2)} \leq \frac{p_{(1,0)}}{q_1(1-q_2)}$
- $\frac{p_{(0,0)}}{(1-q_1)(1-q_2)} \leq \frac{p_{(1,1)}}{q_1 q_2}$

Explicit form of independent weights (cont)

More generally, for each $\omega \in \{0,1\}^d$ and $1 \leq i \leq d$, it is convenient to introduce the auxiliary variables:

$$x_i = \left(\frac{q_i}{1-q_i}\right)^{1-2\omega_i}$$

Explicit form of independent weights (cont)

More generally, for each $\omega \in \{0,1\}^d$ and $1 \leq i \leq d$, it is convenient to introduce the auxiliary variables:

$$x_i = \left(\frac{q_i}{1-q_i}\right)^{1-2\omega_i}$$

Then the problem over each $\mathcal{Q}_\omega$ reduces to:

$$\max_{(x_1,\ldots,x_d)} \quad P(\omega) \cdot \left(1 + \sum_{\alpha \subset \{1,\ldots,d\}} \prod_{i \in \alpha} x_i\right)$$

Explicit form of independent weights (cont)

More generally, for each $\omega \in \{0,1\}^d$ and $1 \leq i \leq d$, it is convenient to introduce the auxiliary variables:

$$x_i = \left(\frac{q_i}{1-q_i}\right)^{1-2\omega_i}$$

Then the problem over each $\mathcal{Q}_\omega$ reduces to:

$$\max_{(x_1,\ldots,x_d)} \quad P(\omega) \cdot \left(1 + \sum_{\alpha \subset \{1,\ldots,d\}} \prod_{i \in \alpha} x_i\right)$$

subject to

$$\prod_{i \in \{1,\ldots,d\} \text{ s.t. } \omega_i' \neq \omega_i} x_i \leq \frac{P(\omega)}{P(\omega')}, \text{ for each } \omega' \neq \omega$$

Explicit form of independent weights (cont)

For example, when $d = 2$, the independent weight $\lambda_{\mathcal{Q}}(P)$ is available as the maximum of the following four quantities:

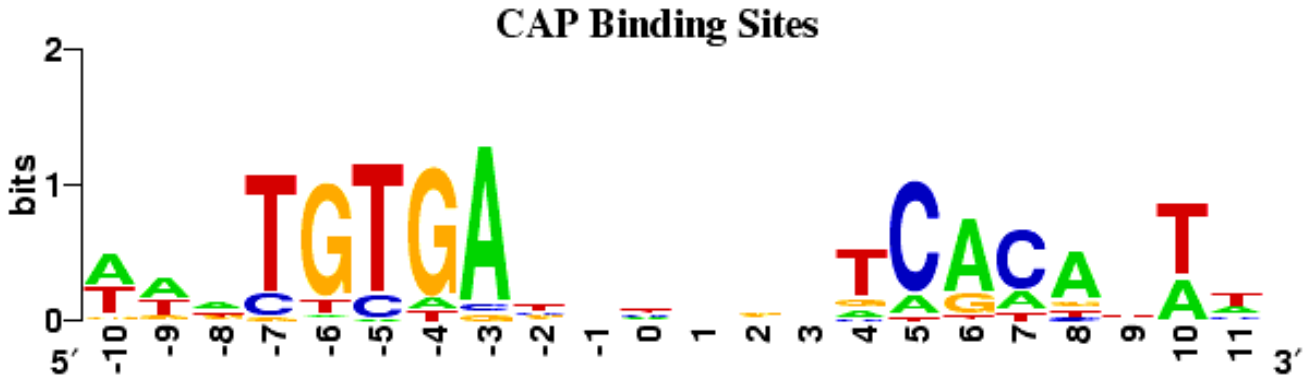
$$\left\{ \begin{array}{ll} p_{10}(1+\frac{p_{00}}{p_{10}})(1+\frac{p_{11}}{p_{10}}) & \text{if } \frac{p_{11}}{p_{10}} \leq \frac{p_{01}}{p_{00}} \\ p_{10}(1+\frac{p_{00}}{p_{10}})(1+\frac{p_{01}}{p_{00}}) \vee p_{10}(1+\frac{p_{01}}{p_{11}})(1+\frac{p_{11}}{p_{10}}) & \text{if } \frac{p_{11}}{p_{10}} > \frac{p_{01}}{p_{00}} \end{array} \right\},$$

$$\left\{ \begin{array}{ll} p_{01}(1+\frac{p_{00}}{p_{01}})(1+\frac{p_{11}}{p_{01}}) & \text{if } \frac{p_{00}}{p_{01}} \leq \frac{p_{10}}{p_{11}} \\ p_{01}(1+\frac{p_{11}}{p_{01}})(1+\frac{p_{10}}{p_{11}}) \vee p_{01}(1+\frac{p_{10}}{p_{00}})(1+\frac{p_{00}}{p_{01}}) & \text{if } \frac{p_{00}}{p_{01}} > \frac{p_{10}}{p_{11}} \end{array} \right\},$$

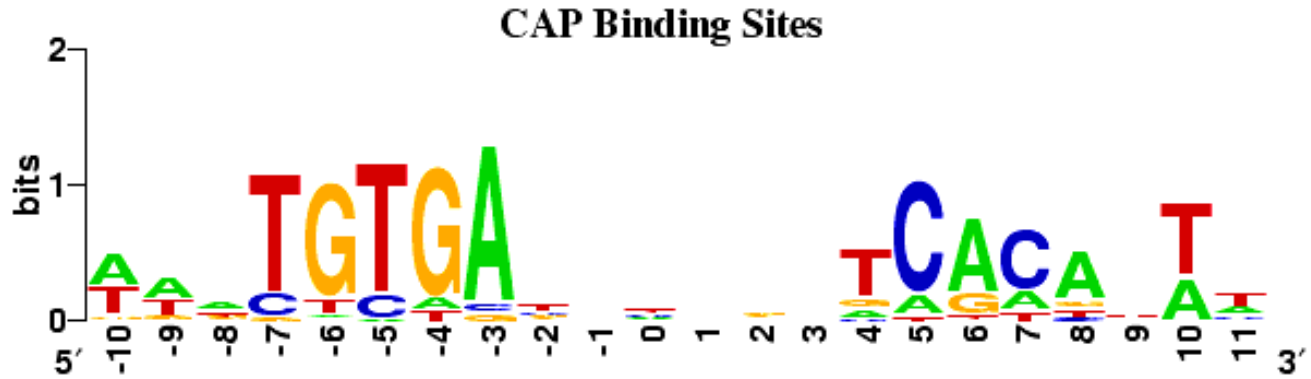
$$\left\{ \begin{array}{ll} p_{00}(1+\frac{p_{10}}{p_{00}})(1+\frac{p_{01}}{p_{00}}) & \text{if } \frac{p_{01}}{p_{00}} \leq \frac{p_{11}}{p_{10}} \\ p_{00}(2+\frac{p_{10}}{p_{00}})(1+\frac{p_{11}}{p_{10}}) \vee p_{00}(1+\frac{p_{11}}{p_{01}})(1+\frac{p_{01}}{p_{00}}) & \text{if } \frac{p_{01}}{p_{00}} > \frac{p_{11}}{p_{10}} \end{array} \right\},$$

$$\left\{ \begin{array}{ll} p_{11}(1+\frac{p_{01}}{p_{11}})(1+\frac{p_{10}}{p_{11}}) & \text{if } \frac{p_{10}}{p_{11}} \leq \frac{p_{00}}{p_{01}} \\ p_{11}(2+\frac{p_{01}}{p_{11}})(1+\frac{p_{00}}{p_{01}}) \vee p_{11}(1+\frac{p_{00}}{p_{10}})(1+\frac{p_{10}}{p_{11}}) & \text{if } \frac{p_{10}}{p_{11}} > \frac{p_{00}}{p_{01}} \end{array} \right\}$$

TFBS data and logo plots



TFBS data and logo plots



Logo plots make the implicit assumption that a nucleotide at one position is independent of the nucleotide at another.

This seems to be approximately true for some transcription factors and false for others.

Can we tell which?

Coding section 4

The highly-dependent short-target TFBS

Logo plot of GATA2, transcription factor known to regulate expression of some embryonic genes

The highly-dependent short-target TFBS

Logo plot of GATA2, transcription factor known to regulate expression of some embryonic genes

"GATA2 binds to a specific nucleic acid sequence viz., (T/A(GATA)A/G), on the promoter and enhancer sites of its target genes, stimulating or suppressing the expression of these target genes. However, there are thousands of sites in human DNA with this nucleotide sequence but for unknown reasons GATA2 binds to $< 1\%$ of these."

Another well-structured class: i.i.d. distributions

Oftentimes researchers assume that a joint distribution of $d > 1$ random variables has independent and identically distributed marginals (i.i.d.-ness)

Another well-structured class: i.i.d. distributions

Oftentimes researchers assume that a joint distribution of $d > 1$ random variables has independent and identically distributed marginals (i.i.d.-ness)

In some cases it is manageable to approximate the weight of the i.i.d. distributions in a model P using a grid. If each variable can take k values, then approximating using a 1% grid requires $\propto 10^{2k-2}$ evaluations of k^d probability ratios $P(\omega)/Q(\omega)$

The i.i.d. weight

Luckily, the i.i.d. weight can be found efficiently in a more explicit form using the previous identity:

$$\lambda_{\mathcal{Q}}(P) = \max_{\omega \in \Omega} \left\{ \max_{Q \in \mathcal{Q}_\omega} \frac{P(\omega)}{Q(\omega)} \right\}$$

The i.i.d. weight

Luckily, the i.i.d. weight can be found efficiently in a more explicit form using the previous identity:

$$\lambda_{\mathcal{Q}}(P) = \max_{\omega \in \Omega} \left\{ \max_{Q \in \mathcal{Q}_\omega} \frac{P(\omega)}{Q(\omega)} \right\}$$

Suppose each random variable is binary. Define $\tilde{P}(\omega)$ to be the minimum $P(\omega')$, where ω' is any rearrangement of ω. Then

$$\lambda_{\mathcal{Q}}(P) = \max_{\omega \in \{0,1\}^d} \left\{ \max_{Q \in \mathcal{Q}_\omega \leftrightarrow q} \frac{\tilde{P}(\omega)}{q^{\#\{1 \in \omega\}}(1-q)^{d - \#\{1 \in \omega\}}} \right\}$$

The i.i.d. weight (cont)

After thoughtful consideration the reparameterization $r := \frac{1-q}{q}$ allows us to derive a closed form of the maximum in each $\mathcal{Q}_\omega$:

$$\max_{q \in \mathcal{Q}_\omega} \frac{\tilde{P}(\omega)}{q^{\#\{1 \in \omega\}}(1-q)^{d-\#\{1 \in \omega\}}} = \tilde{P}(\omega) \cdot \max_{r \geq 0} \frac{(1+r)^d}{r^{d-\#\{1 \in \omega\}}}$$

The i.i.d. weight (cont)

After thoughtful consideration the reparameterization $r := \frac{1-q}{q}$ allows us to derive a closed form of the maximum in each $\mathcal{Q}_\omega$:

$$\max_{q \in \mathcal{Q}_\omega} \frac{\tilde{P}(\omega)}{q^{\#\{1 \in \omega\}}(1-q)^{d-\#\{1 \in \omega\}}} = \tilde{P}(\omega) \cdot \max_{r \geq 0} \frac{(1+r)^d}{r^{d-\#\{1 \in \omega\}}}$$

subject to: $r^{\#\{1 \in \omega\} - \#\{1 \in \omega'\}} \leq \frac{\tilde{P}(\omega)}{\tilde{P}(\omega')}$ for each $\omega' \in \{0,1\}^d$

The i.i.d. weight (cont)

After thoughtful consideration the reparameterization $r := \frac{1-q}{q}$ allows us to derive a closed form of the maximum in each $\mathcal{Q}_\omega$:

$$\max_{q \in \mathcal{Q}_\omega} \frac{\tilde{P}(\omega)}{q^{\#\{1 \in \omega\}}(1-q)^{d-\#\{1 \in \omega\}}} = \tilde{P}(\omega) \cdot \max_{r \geq 0} \frac{(1+r)^d}{r^{d-\#\{1 \in \omega\}}}$$

subject to: $\quad r^{\#\{1 \in \omega\} - \#\{1 \in \omega'\}} \leq \frac{\tilde{P}(\omega)}{\tilde{P}(\omega')}$ for each $\omega' \in \{0,1\}^d$

The constraints on r are equivalent to:

$$r \in \Bigg[\max_{\{\alpha \in \{0,1\}^d \mid \#\{1 \in \alpha\} > \#\{1 \in \omega\}\}} \left\{ \left(\frac{\tilde{P}(\omega)}{\tilde{P}(\alpha)} \right)^{\frac{1}{\#\{1 \in \alpha\} - \#\{1 \in \omega\}}} \right\}, \min_{\{\beta \in \{0,1\}^d \mid \#\{1 \in \beta\} < \#\{1 \in \omega\}\}} \left\{ \left(\frac{\tilde{P}(\omega)}{\tilde{P}(\alpha)} \right)^{\frac{1}{\#\{1 \in \alpha\} - \#\{1 \in \omega\}}} \right\} \Bigg]$$

Further analysis reveals that the max occurs at one of the above endpoints. Computing $\lambda_{\mathcal{Q}}(P)$ therefore requires calculating only $d(d+1)$ quantities!

Another well-structured class: the Poisson weight

Suppose we want to find the latent weight of a model P with sample space $\Omega = \{0, 1, 2, \ldots \ell\}$ w.r.t. the class $\mathcal{Q}$ of truncated Poisson distributions

Another well-structured class: the Poisson weight

Suppose we want to find the latent weight of a model P with sample space $\Omega = \{0, 1, 2, \ldots \ell\}$ w.r.t. the class $\mathcal{Q}$ of truncated Poisson distributions

This requires maximizing:

$$F(\alpha) := \min \left\{ \min_{i=0,\ldots,\ell-1} \frac{P(i)}{\alpha^i e^{-\alpha}/i!}, \frac{1 - \sum_{i=0}^{\ell-1} P(i)}{1 - \sum_{i=0}^{\ell-1} \alpha^i e^{-\alpha}/i!} \right\},$$

over $\alpha > 0$

The non-differentiable function $F(\alpha)$ has various other properties that allow its minimization efficiently!

The Poisson weight of a truncated model

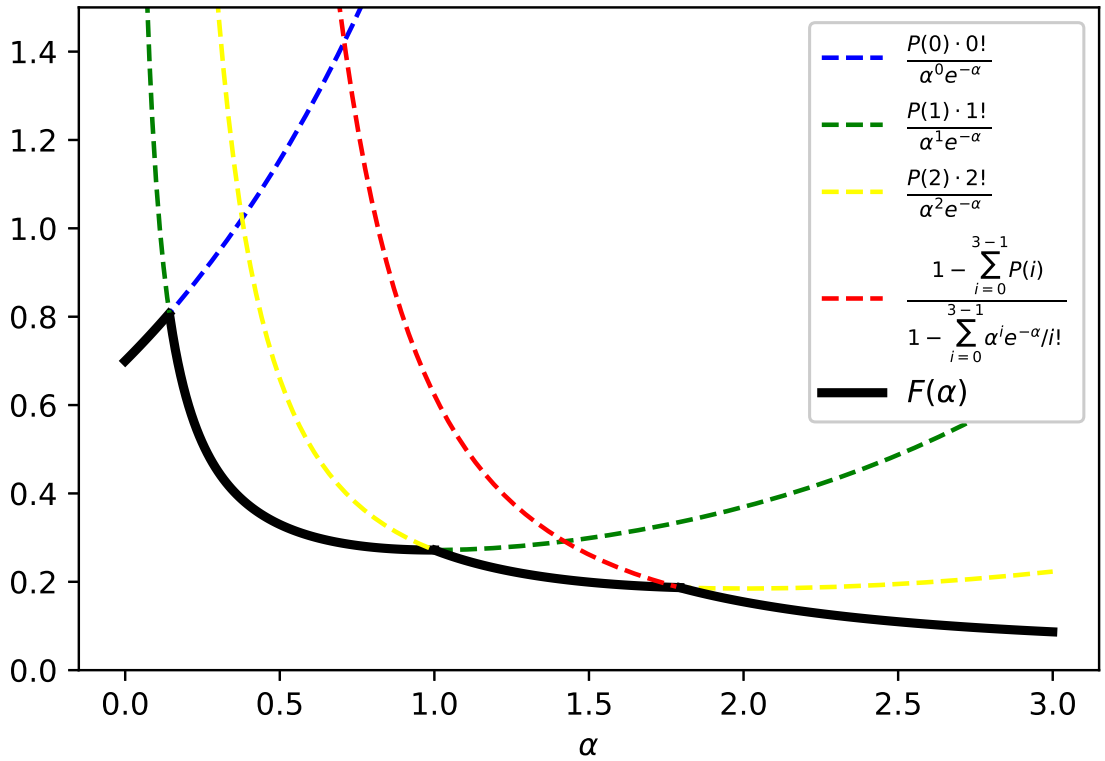


Illustration of how to maximize the function $F(\alpha)$ when $\ell = 3$

Example: Bortkiewicz's horse kick studies

Deaths	0	1	2	3	4	5+	Total
Number	144	91	32	11	2	0	280

Table: Reported number of horse-kick deaths per-year in 14 army corps of the Prussian Army between 1875-1894. No army corp reported five or more deaths in a single year during this period

Coding section 5

Another class $\mathcal{Q}$, the exchangeable models

Random variables $(X_1, \ldots, X_d)$ are *exchangeable*, or their joint distribution P is said to be exchangeable, if their probability is invariant to "shuffling," i.e., $(X_{\sigma(1)}, \ldots, X_{\sigma(d)}) \stackrel{d}{=} (X_1, \ldots, X_d)$ for any permutation σ of $\{1, \ldots, d\}$

Another class $\mathcal{Q}$, the exchangeable models

Random variables $(X_1, \ldots, X_d)$ are *exchangeable*, or their joint distribution P is said to be exchangeable, if their probability is invariant to "shuffling," i.e., $(X_{\sigma(1)}, \ldots, X_{\sigma(d)}) \stackrel{d}{=} (X_1, \ldots, X_d)$ for any permutation σ of $\{1, \ldots, d\}$

The set of exchangeable distributions contains all i.i.d. probability models but is generally much larger. It arises from quite natural urn sampling regimes, is the critical hypothesis in DeFinetti's theorem, and is a property of Bayesian nonparametric models and various random graph models.

Exchangeable weight properties

Suppose each random variable $X_1, \ldots, X_d$ takes values in $\{1, \ldots, k\}$ For each outcome $\omega \in \{1, \ldots, k\}^d$, define the *permutation-equivalence class of* ω as

$$[\omega] := \text{the set of outcomes that can be formed by rearranging } \omega$$

(The set of all $[\omega]$ is a partition of the sample space)

Exchangeable weight properties

Suppose each random variable $X_1, \ldots, X_d$ takes values in $\{1, \ldots, k\}$ For each outcome $\omega \in \{1, \ldots, k\}^d$, define the *permutation-equivalence class of* ω as

$$[\omega] := \text{the set of outcomes that can be formed by rearranging } \omega$$

(The set of all $[\omega]$ is a partition of the sample space)

Then

$$\lambda_{\mathcal{Q}}(P) = \sum_{\omega \in \{1,\ldots,k\}^d} \min_{y \in [\omega]} P(y) = \sum_{[\omega] \subset \{1,\ldots,k\}^d} |[\omega]| \cdot \min_{y \in [\omega]} P(y).$$

Moreover, the unique exchangeable distribution Q that achieves weight $\lambda_{\mathcal{Q}}(P)$ as a component of P gives mass $\min_{y \in [\omega]} P(y)/\lambda_{\mathcal{Q}}(P)$ to each outcome ω

Exchangeable weight properties (cont)

Note that in most cases

$$P = \lambda(P) \cdot Q + (1 - \lambda) \cdot R,$$

with a unique Q and R. It can also be shown that $\lambda_{\mathcal{Q}}(R) = 0$

Exchangeable weight properties (cont)

Note that in most cases

$$P = \lambda(P) \cdot Q + (1 - \lambda) \cdot R,$$

with a unique Q and R. It can also be shown that $\lambda_{\mathcal{Q}}(R) = 0$

In this sense, data from P can be distilled entirely into *exchangeable* and *unexchangeable* parts

Exchangeable weight properties (cont)

The number of permutation-equivalence classes is $\binom{k+d-1}{d}$, which may be very large. For this reason, we may wish to compute an easier approximation of $\lambda_{\mathcal{Q}}(P)$.

Exchangeable weight properties (cont)

The number of permutation-equivalence classes is $\binom{k+d-1}{d}$, which may be very large. For this reason, we may wish to compute an easier approximation of $\lambda_{\mathcal{Q}}(P)$.

- If $S \in \mathcal{P}(\{1, \ldots, k\}^{d-1})$ is a joint marginal of P, then $\lambda_{\mathcal{Q}}(P) \leq \lambda_{\mathcal{Q}}(S)$.

Exchangeable weight properties (cont)

The number of permutation-equivalence classes is $\binom{k+d-1}{d}$, which may be very large. For this reason, we may wish to compute an easier approximation of $\lambda_{\mathcal{Q}}(P)$.

- If $S \in \mathcal{P}(\{1, \ldots, k\}^{d-1})$ is a joint marginal of P, then $\lambda_{\mathcal{Q}}(P) \leq \lambda_{\mathcal{Q}}(S)$.
- If $\tilde{P}$ is a "lumped" version of P then $\lambda_{\mathcal{Q}}(P) \leq \lambda_{\mathcal{Q}}(\tilde{P})$.

Exchangeable weight properties (cont)

The number of permutation-equivalence classes is $\binom{k+d-1}{d}$, which may be very large. For this reason, we may wish to compute an easier approximation of $\lambda_{\mathcal{Q}}(P)$.

- If $S \in \mathcal{P}(\{1, \ldots, k\}^{d-1})$ is a joint marginal of P, then $\lambda_{\mathcal{Q}}(P) \leq \lambda_{\mathcal{Q}}(S)$.
- If $\tilde{P}$ is a "lumped" version of P then $\lambda_{\mathcal{Q}}(P) \leq \lambda_{\mathcal{Q}}(\tilde{P})$.

Both properties also apply to latent weights of the independent or i.i.d. models

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$
- Because $\lambda_{\mathcal{Q}}$ is continuous, $\hat{\lambda}_n \overset{a.s.}{\to} \lambda_{\mathcal{Q}}(P)$.

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$
- Because $\lambda_{\mathcal{Q}}$ is continuous, $\hat{\lambda}_n \overset{a.s.}{\to} \lambda_{\mathcal{Q}}(P)$.
- Under certain regularity conditions the multivariate delta method applies and $\hat{\lambda}_n$ has an easily-computed limiting Gaussian distribution

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$
- Because $\lambda_{\mathcal{Q}}$ is continuous, $\hat{\lambda}_n \overset{a.s.}{\to} \lambda_{\mathcal{Q}}(P)$.
- Under certain regularity conditions the multivariate delta method applies and $\hat{\lambda}_n$ has an easily-computed limiting Gaussian distribution
  - For large n, $\hat{\lambda}_n \overset{d}{\approx} N\Big(\lambda_{\mathcal{Q}}(P), \nabla^t \lambda_{\mathcal{Q}}(P) \cdot \Sigma \cdot \nabla \lambda_{\mathcal{Q}}(P)\Big)$, where Σ contains $p_i(1-p_i)/n$ on the diagonal and $-p_i p_j/n$ elsewhere. Usefully, $\nabla \lambda_{\mathcal{Q}}(P)$ is piecewise constant

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$
- Because $\lambda_{\mathcal{Q}}$ is continuous, $\hat{\lambda}_n \overset{a.s.}{\to} \lambda_{\mathcal{Q}}(P)$.
- Under certain regularity conditions the multivariate delta method applies and $\hat{\lambda}_n$ has an easily-computed limiting Gaussian distribution
  - For large n, $\hat{\lambda}_n \overset{d}{\approx} N\Big(\lambda_{\mathcal{Q}}(P), \nabla^t \lambda_{\mathcal{Q}}(P) \cdot \Sigma \cdot \nabla \lambda_{\mathcal{Q}}(P)\Big)$, where Σ contains $p_i(1-p_i)/n$ on the diagonal and $-p_i p_j/n$ elsewhere. Usefully, $\nabla \lambda_{\mathcal{Q}}(P)$ is piecewise constant
- The bias $E(\hat{\lambda}_n - \lambda_{\mathcal{Q}}(P)) \leq 0$, and $\hat{\lambda}_n$ is asymptotically unbiased

Estimating the exchangeable weight

Suppose $\mathbb{X}_1, \ldots, \mathbb{X}_n$ is an i.i.d. sample from an unknown source P

- Evaluating $\lambda_{\mathcal{Q}}$ at the empirical distribution $\hat{P}_n = \sum_{i=1}^{n} \delta_{\mathbb{X}_i}/n$, $\hat{\lambda}_n := \lambda_{\mathcal{Q}}(\hat{P}_n)$ is a MLE for $\lambda_{\mathcal{Q}}(P)$
- Because $\lambda_{\mathcal{Q}}$ is continuous, $\hat{\lambda}_n \overset{a.s.}{\to} \lambda_{\mathcal{Q}}(P)$.
- Under certain regularity conditions the multivariate delta method applies and $\hat{\lambda}_n$ has an easily-computed limiting Gaussian distribution
  - For large n, $\hat{\lambda}_n \overset{d}{\approx} N\Big(\lambda_{\mathcal{Q}}(P), \nabla^t \lambda_{\mathcal{Q}}(P) \cdot \Sigma \cdot \nabla \lambda_{\mathcal{Q}}(P)\Big)$, where Σ contains $p_i(1-p_i)/n$ on the diagonal and $-p_i p_j/n$ elsewhere. Usefully, $\nabla \lambda_{\mathcal{Q}}(P)$ is piecewise constant
- The bias $E(\hat{\lambda}_n - \lambda_{\mathcal{Q}}(P)) \leq 0$, and $\hat{\lambda}_n$ is asymptotically unbiased
- Under regularity conditions, the bootstrap estimator is consistent

A quick aside about estimating latent weights

If a source P (over any sample space) is observed only indirectly through data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, the sampling distribution $\lambda_{\mathcal{Q}}(\hat{P}_n)$ may be rather opaque

A quick aside about estimating latent weights

If a source P (over any sample space) is observed only indirectly through data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, the sampling distribution $\lambda_{\mathcal{Q}}(\hat{P}_n)$ may be rather opaque

If an explicit or semi-explicit form of the latent weight is known (such as when $\mathcal{Q}$ is the set of Poisson distributions, as above), we may be able to bound the sampling error $|\lambda_{\mathcal{Q}}(\hat{P}_n) - \lambda_{\mathcal{Q}}(P)|$ with specified probability.

A quick aside about estimating latent weights

If a source P (over any sample space) is observed only indirectly through data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, the sampling distribution $\lambda_{\mathcal{Q}}(\hat{P}_n)$ may be rather opaque

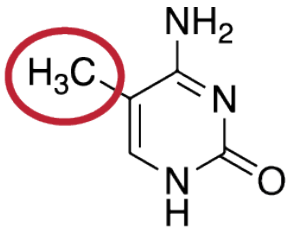
If an explicit or semi-explicit form of the latent weight is known (such as when $\mathcal{Q}$ is the set of Poisson distributions, as above), we may be able to bound the sampling error $|\lambda_{\mathcal{Q}}(\hat{P}_n) - \lambda_{\mathcal{Q}}(P)|$ with specified probability.

However, in most cases the sampling distribution should not be assumed to be asymptotically normal. The asymptotic sampling distribution is often a somewhat complicated function of a multivariate normal random vector.

Is methylation exchangeable over short stretches of DNA?

- Cytosines in a "CpG" context may be methylated or unmethylated

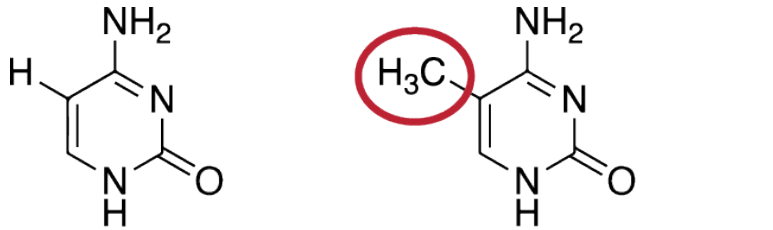
Cytosine



methylated Cytosine

Is methylation exchangeable over short stretches of DNA?

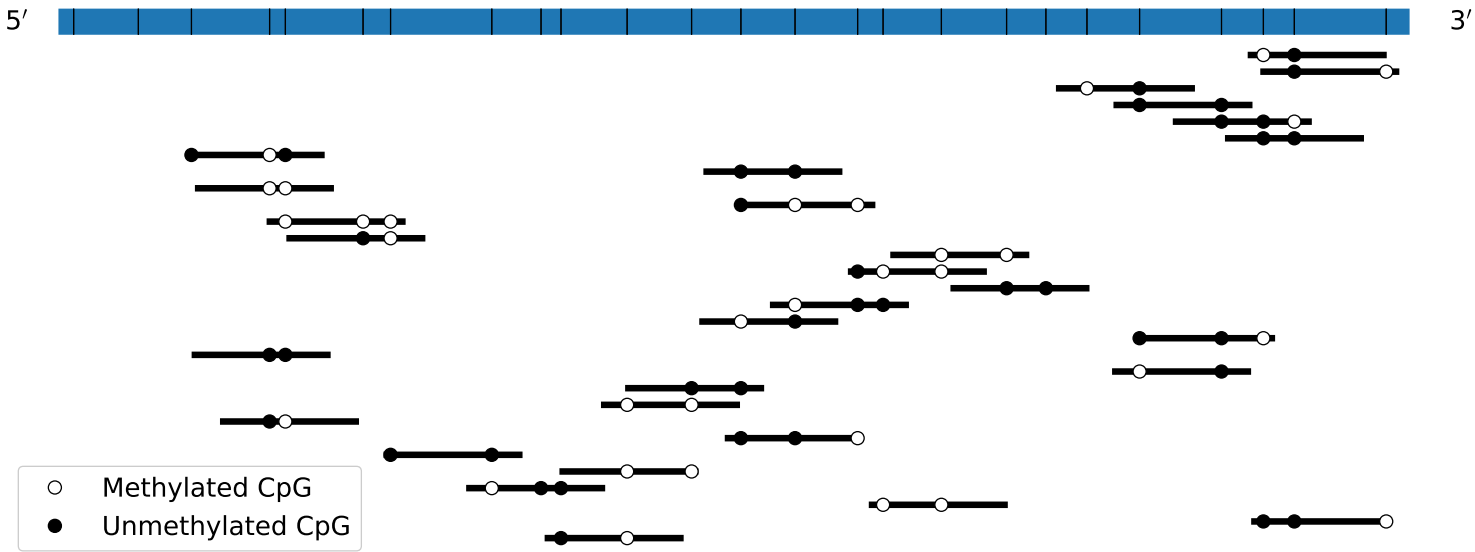
- Cytosines in a "CpG" context may be methylated or unmethylated



Cytosine **methylated Cytosine**

- The sequence of CpGs can be modeled as a random binary sequence

Is methylation exchangeable over short stretches of DNA? (cont)



A typical whole-genome bisulfite sequencing (WGBS) experiment.

Is methylation exchangeable over short stretches of DNA? (cont)

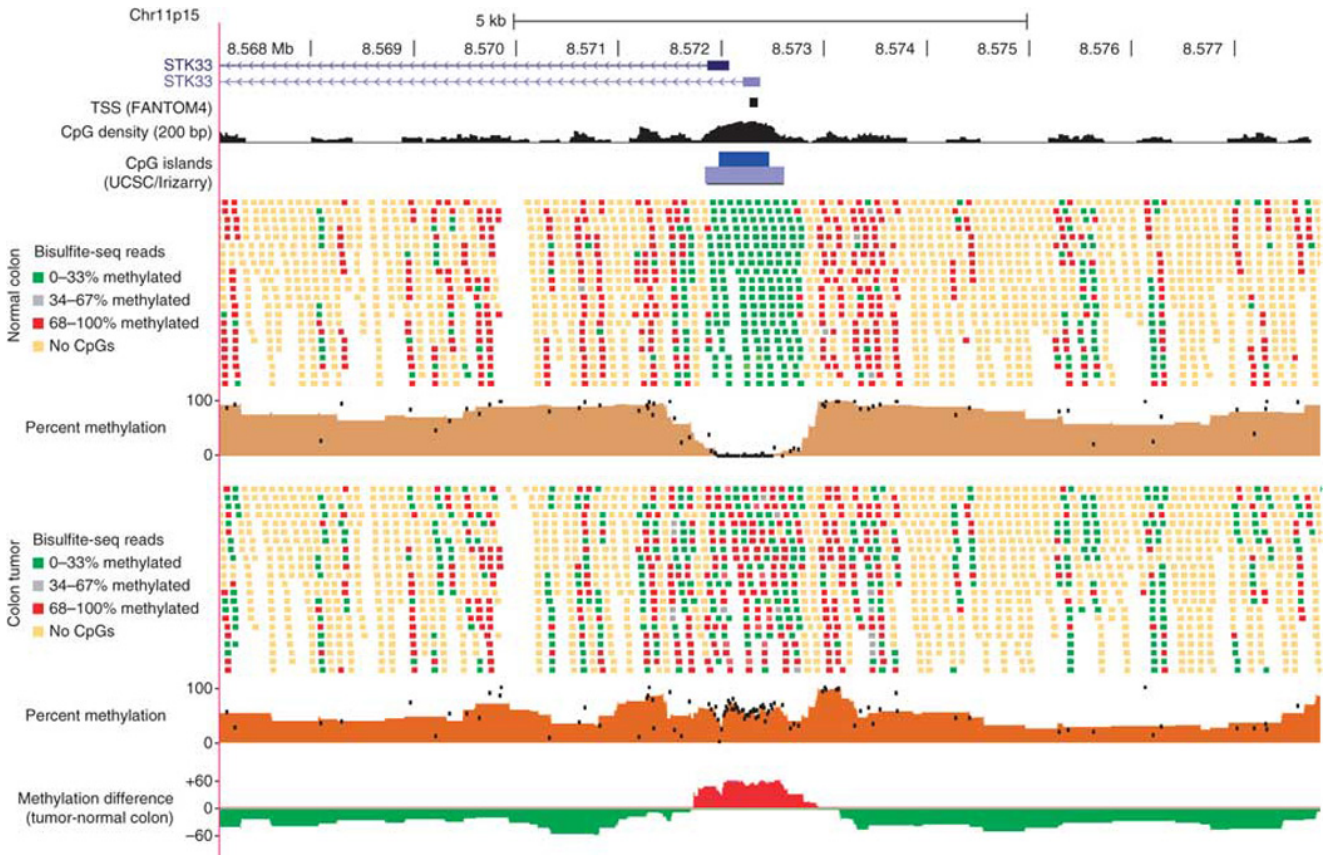
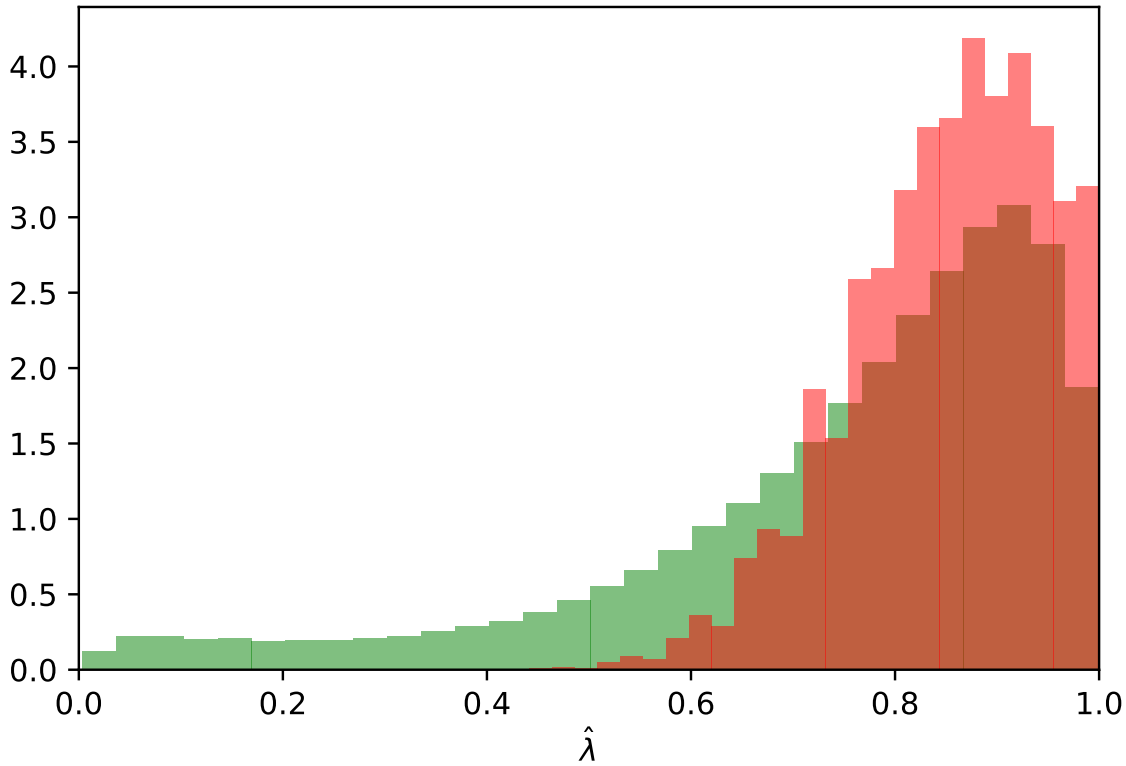


Figure: Berman et al., 2012; sliding window approach

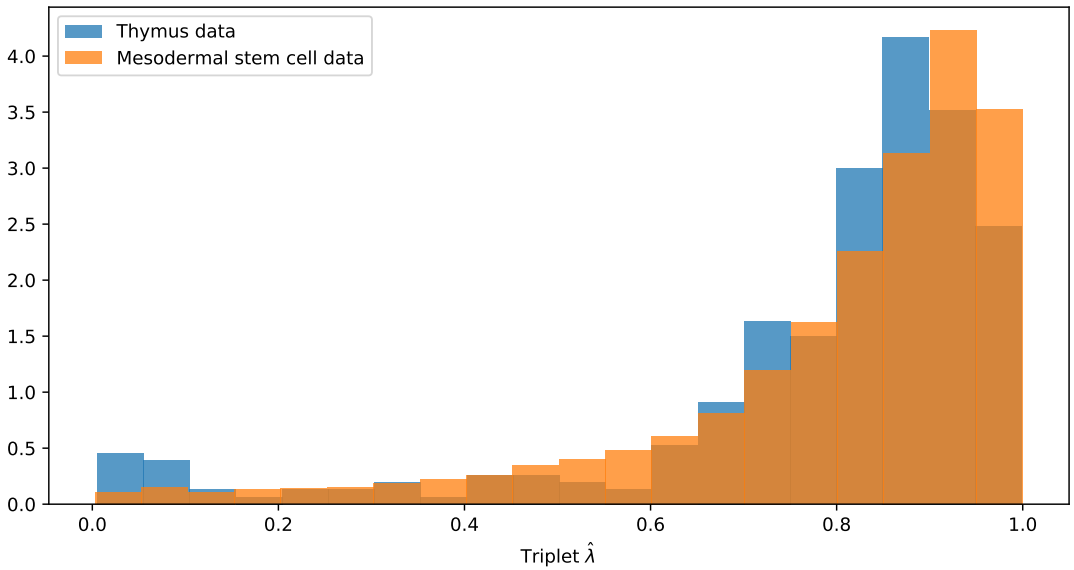
Coding section 6

Is methylation exchangeable over short stretches of DNA? (cont)



Negative bias or sampling error of exchangeable models (red, synthetic worst-case data with $n = 100$) does not adequately explain the appearance of so many un-exchangeable loci in real WGBS data (green).

Is methylation exchangeable over short stretches of DNA? (cont)



Final remarks

- Latent weights offer a different perspective on the structure contained inside a probabilistic source

Final remarks

- Latent weights offer a different perspective on the structure contained inside a probabilistic source
- When a source P is known exactly it is often possible to numerically approximate the latent weight $\lambda_{\mathcal{Q}}(P)$ of a class $\mathcal{Q}$ by placing a numerical grid over $\mathcal{Q}$

Final remarks

- Latent weights offer a different perspective on the structure contained inside a probabilistic source
- When a source P is known exactly it is often possible to numerically approximate the latent weight $\lambda_{\mathcal{Q}}(P)$ of a class $\mathcal{Q}$ by placing a numerical grid over $\mathcal{Q}$
- When $\mathcal{Q}$ describes a family of well-structured models it is often possible to find an explicit or semi-explicit form of $\lambda_{\mathcal{Q}}(P)$

Final remarks

- Latent weights offer a different perspective on the structure contained inside a probabilistic source
- When a source P is known exactly it is often possible to numerically approximate the latent weight $\lambda_{\mathcal{Q}}(P)$ of a class $\mathcal{Q}$ by placing a numerical grid over $\mathcal{Q}$
- When $\mathcal{Q}$ describes a family of well-structured models it is often possible to find an explicit or semi-explicit form of $\lambda_{\mathcal{Q}}(P)$
- These more explicit forms are useful in determining the sampling distribution when P is observed only indirectly through a random sample

Final remarks

- Latent weights offer a different perspective on the structure contained inside a probabilistic source
- When a source P is known exactly it is often possible to numerically approximate the latent weight $\lambda_{\mathcal{Q}}(P)$ of a class $\mathcal{Q}$ by placing a numerical grid over $\mathcal{Q}$
- When $\mathcal{Q}$ describes a family of well-structured models it is often possible to find an explicit or semi-explicit form of $\lambda_{\mathcal{Q}}(P)$
- These more explicit forms are useful in determining the sampling distribution when P is observed only indirectly through a random sample
- With very large data a latent weight may give us information on when it is okay to model data from P as having $\mathcal{Q}$'s special properties, even if we have rejected the null hypothesis that $P \in \mathcal{Q}$

Acknowledgements

- Manuel Lladser (advisor)
- Francis Baffour (collaborator)

This work has been partially funded by the NSF GRFP grant No. 2016198773 (Pearson), and the NSF IGERT grant No. 1144807 (BioFrontiers Institute/IQ Biology Program)