Lecture 6. June 14 2019

Recap

The last lecture gave an overview of the cancer world, describing some of the biological background and some of the opportunities for statisticians, computational biologists, machine learners and data scientists

Recap

The last lecture gave an overview of the cancer world, describing some of the biological background and some of the opportunities for statisticians, computational biologists, machine learners and data scientists

It also described the new Morningside Heights *Irving Institute for Cancer Dynamics* – its website is given below

Recap

The last lecture gave an overview of the cancer world, describing some of the biological background and some of the opportunities for statisticians, computational biologists, machine learners and data scientists

It also described the new Morningside Heights *Irving Institute for Cancer Dynamics* – its website is given below

Reminder: the slides, and handout versions, are available at `https://cancerdynamics.columbia.edu/content/summer-program`

Recap

The last lecture gave an overview of the cancer world, describing some of the biological background and some of the opportunities for statisticians, computational biologists, machine learners and data scientists

It also described the new Morningside Heights *Irving Institute for Cancer Dynamics* – its website is given below

Reminder: the slides, and handout versions, are available at `https://cancerdynamics.columbia.edu/content/summer-program`

Today it is back to ABC, specifically about how to find summary statistics

Summary Statistics

Reference: Prangle D (2018) Chapter 5 in *Handbook of Approximate Bayesian Computation*, CRC Press

We have seen that to deal with high-dimensional data we reduce them to lower dimensional summary statistics, and make comparisons between the summary statistics of the simulation and the observed data

Summary Statistics

Reference: Prangle D (2018) Chapter 5 in *Handbook of Approximate Bayesian Computation*, CRC Press

We have seen that to deal with high-dimensional data we reduce them to lower dimensional summary statistics, and make comparisons between the summary statistics of the simulation and the observed data

We have a version of the curse of dimensionality, in this case the number of summary statistics used: too many makes approximation to posterior worse, too few might lose important features of the data.

Aim: balance *low dimension* and *informativeness*

Summary Statistics

Reference: Prangle D (2018) Chapter 5 in *Handbook of Approximate Bayesian Computation*, CRC Press

We have seen that to deal with high-dimensional data we reduce them to lower dimensional summary statistics, and make comparisons between the summary statistics of the simulation and the observed data

We have a version of the curse of dimensionality, in this case the number of summary statistics used: too many makes approximation to posterior worse, too few might lose important features of the data.

Aim: balance *low dimension* and *informativeness*

Note: we focus today on continuous parameters (discrete parameters use similar techniques to the model choice setting)

The curse of dimensionality

There are some theoretical results in the literature, for example for MSE of a standard ABC rejection sampling method (see Barber et al. *Elec J Stats*, **9**, 80–105) being of the form

$$O_p\left(n^{-4/(q+4)}\right)$$

where $q = \dim(S(\mathcal{D}))$

The curse of dimensionality

There are some theoretical results in the literature, for example for MSE of a standard ABC rejection sampling method (see Barber et al. *Elec J Stats*, **9**, 80–105) being of the form

$$O_p\left(n^{-4/(q+4)}\right)$$

where $q = \dim(S(\mathcal{D}))$

The warning is that for larger ϵ, high dimensional summaries typically give poor results

The curse of dimensionality

There are some theoretical results in the literature, for example for MSE of a standard ABC rejection sampling method (see Barber et al. *Elec J Stats*, **9**, 80–105) being of the form

$$O_p\left(n^{-4/(q+4)}\right)$$

where $q = \dim(S(\mathcal{D}))$

The warning is that for larger ϵ, high dimensional summaries typically give poor results

Can get some leverage if likelihood factorises (as in the primate example), where can use ABC in each factor (which is of lower dimension)

Sufficiency

Sufficiency

S is sufficient for θ if $f(\mathcal{D}|S,\theta)$ does not depend on θ

Sufficiency

Sufficiency

S is sufficient for θ if $f(\mathcal{D}|S,\theta)$ does not depend on θ

Bayes sufficiency

S is Bayes sufficient for θ if $\theta|S$ and $\theta|\mathcal{D}$ have same distribution for any prior and almost all $\mathcal{D}$

Sufficiency

Sufficiency

S is sufficient for θ if $f(\mathcal{D}|S,\theta)$ does not depend on θ

Bayes sufficiency

S is Bayes sufficient for θ if $\theta|S$ and $\theta|\mathcal{D}$ have same distribution for any prior and almost all $\mathcal{D}$

The latter is the natural definition of sufficiency for ABC: an ABC algorithm with Bayes sufficient S and $\epsilon \to 0$ results in the correct posterior

Strategies for selecting summary statistics

Three rough groupings:

- Subset selection
- Projection
- Auxiliary likelihood (*not doing this today*)

Strategies for selecting summary statistics

Three rough groupings:

- Subset selection
- Projection
- Auxiliary likelihood (*not doing this today*)

The first two methods start with choice of a set of data features, $Z = Z(\mathcal{D})$.

- For subset selection, these are candidate summary statistics
- Both methods need training data $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n_0$

Strategies for selecting summary statistics

Three rough groupings:

- Subset selection
- Projection
- Auxiliary likelihood (*not doing this today*)

The first two methods start with choice of a set of data features, $Z = Z(\mathcal{D})$.

- For subset selection, these are candidate summary statistics
- Both methods need training data $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n_0$
- Subset selection methods choose a subset of Z, by optimizing some criterion on a training set
- Projection methods use training set to choose a projection of Z, resulting in dimension reduction

Auxiliary likelihood methods:

- Do not need a feature set or a training set

Auxiliary likelihood methods:

- Do not need a feature set or a training set
- Rather, they exploit an approximating model whose likelihood (the auxiliary likelihood) is more tractable than the model of interest

Auxiliary likelihood methods:

- Do not need a feature set or a training set
- Rather, they exploit an approximating model whose likelihood (the auxiliary likelihood) is more tractable than the model of interest
- Composite likelihood used as an approximation

Auxiliary likelihood methods:

- Do not need a feature set or a training set
- Rather, they exploit an approximating model whose likelihood (the auxiliary likelihood) is more tractable than the model of interest
- Composite likelihood used as an approximation
- Often exploit subject area knowledge (as in our population genetics problems)

Auxiliary likelihood methods:

- Do not need a feature set or a training set
- Rather, they exploit an approximating model whose likelihood (the auxiliary likelihood) is more tractable than the model of interest
- Composite likelihood used as an approximation
- Often exploit subject area knowledge (as in our population genetics problems)
- Derives summaries from the simpler model

Subset selection

A variety of approaches:

1. Joyce and Marjoram (2008) Approximate sufficiency

- Idea: if S is sufficient, then the posterior distribution for θ will be unaffected by replacing S with $S' = S \cup X$, where X is an additional summary statistic

Subset selection

A variety of approaches:

1. Joyce and Marjoram (2008) Approximate sufficiency

- Idea: if S is sufficient, then the posterior distribution for θ will be unaffected by replacing S with $S' = S \cup X$, where X is an additional summary statistic
- Works for one-dimensional θ. Not clear how to implement in higher dimensions

Subset selection

A variety of approaches:

1. Joyce and Marjoram (2008) Approximate sufficiency

- Idea: if S is sufficient, then the posterior distribution for θ will be unaffected by replacing S with $S' = S \cup X$, where X is an additional summary statistic
- Works for one-dimensional θ. Not clear how to implement in higher dimensions

2. Nunes and Balding (2010) Entropy/loss minimization

- Start with a universe of summaries, $S \subset \Omega$
- Generate parameter values and datasets $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n$

Subset selection

A variety of approaches:

1. Joyce and Marjoram (2008) Approximate sufficiency

- Idea: if S is sufficient, then the posterior distribution for θ will be unaffected by replacing S with $S' = S \cup X$, where X is an additional summary statistic
- Works for one-dimensional θ. Not clear how to implement in higher dimensions

2. Nunes and Balding (2010) Entropy/loss minimization

- Start with a universe of summaries, $S \subset \Omega$
- Generate parameter values and datasets $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n$

  Rejection-ABC:

  - Compute the values of S, say S_i, for ith data set, and accept the θ_i corresponding to the n_0 smallest values of $||S_i - S^*||$, where S^* is the value of S for the observed data.

- ME:
  - For each $S \subset \Omega$, do Rejection-ABC and compute $\hat{H}$ from the n_0 accepted values. S_{ME} is the value of S that minimizes $\hat{H}$, and the corresponding values of θ_i give the approximation to the posterior for θ

- ME:
  - For each $S \subset \Omega$, do Rejection-ABC and compute $\hat{H}$ from the n_0 accepted values. S_{ME} is the value of S that minimizes $\hat{H}$, and the corresponding values of θ_i give the approximation to the posterior for θ

Here, $\hat{H}$ is the kth nearest neighbor estimator of entropy, given by

$$\hat{H} = \log\left(\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right) - \psi(k) + \log n + \frac{p}{n}\sum_{i=1}^{n} \log R_{ik};$$

p is the dimension of θ, R_{ik} is the Euclidean distance from θ_i to its kth nearest neighbour in the posterior sample, and $\psi(x) = \Gamma'(x)/\Gamma(x)$.

For large Ω, rather than consider all subsets of Ω it might be necessary to restrict attention to subsets such as $\{S \subset \Omega : |S| < k\}$, or some investigator specified set

For large Ω, rather than consider all subsets of Ω it might be necessary to restrict attention to subsets such as $\{S \subset \Omega : |S| < k\}$, or some investigator specified set

Nunes and Balding suggest a two-stage version as well. This aims to find the subset of Z that optimizes the performance of ABC on datasets similar to $\mathcal{D}_{\text{obs}}$, by minimising the average of the RMSE loss function

$$\left(\frac{1}{t} \sum_{i=1}^{t} ||\theta_i - \theta'||_2 \right)^{1/2}$$

Here θ' is the parameter value that generated $\mathcal{D}'$, and $(\theta_i, 1 \leq i \leq t)$ is the ABC output sample when $\mathcal{D}'$ is used as the observations.

For large Ω, rather than consider all subsets of Ω it might be necessary to restrict attention to subsets such as $\{S \subset \Omega : |S| < k\}$, or some investigator specified set

Nunes and Balding suggest a two-stage version as well. This aims to find the subset of Z that optimizes the performance of ABC on datasets similar to $\mathcal{D}_{\text{obs}}$, by minimising the average of the RMSE loss function

$$\left(\frac{1}{t} \sum_{i=1}^{t} ||\theta_i - \theta'||_2 \right)^{1/2}$$

Here θ' is the parameter value that generated $\mathcal{D}'$, and $(\theta_i, 1 \leq i \leq t)$ is the ABC output sample when $\mathcal{D}'$ is used as the observations.

They suggest selecting summaries by entropy minimization and perform ABC to generate $(\theta', \mathcal{D}')$ pairs. Then select summaries by minimising RMSE.

3. Barnes et al. (2012), Filippi et al. (2012) Mutual information

Sufficiency can be restated in terms of mutual information: sufficient statistics maximise the mutual information between $S(\mathcal{D})$ and θ. A necessary condition for sufficiency is that the KL divergence of $f(\theta|S(\mathcal{D}))$ from $f(\theta|\mathcal{D})$ is zero:

$$\int f(\theta|\mathcal{D}) \log \left(\frac{f(\theta|\mathcal{D})}{f(\theta|S(\mathcal{D}))} \right) = 0.$$

3. Barnes et al. (2012), Filippi et al. (2012) Mutual information

Sufficiency can be restated in terms of mutual information: sufficient statistics maximise the mutual information between $S(\mathcal{D})$ and θ. A necessary condition for sufficiency is that the KL divergence of $f(\theta|S(\mathcal{D}))$ from $f(\theta|\mathcal{D})$ is zero:

$$\int f(\theta|\mathcal{D}) \log \left(\frac{f(\theta|\mathcal{D})}{f(\theta|S(\mathcal{D}))} \right) = 0.$$

This suggests a stepwise selection method to choose a subset of Z:

- Add a new statistic z to existing subset S, creating a new subset S', if the estimated KL divergence of $f_{\mathrm{ABC}}(\theta|S(\mathcal{D}))$ from $f_{\mathrm{ABC}}(\theta|S'(\mathcal{D}))$ is above some threshold; here $\mathcal{D}$ is the observed data.

4. Sedki and Pudlo (2012), Blum et al. (2013) Regularisation approaches

The idea is to

- fit a linear regression with response θ and covariates Z based on training data
- use variable selection to find an informative subset of Z (e.g. via minimising AIC, or BIC)

Might also use lasso to simplify the computation.

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard
- Regularization methods cheaper, but less studied and so less well understood

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard
- Regularization methods cheaper, but less studied and so less well understood
- Interpretable output

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard
- Regularization methods cheaper, but less studied and so less well understood
- Interpretable output
- Useful for model criticism: if some marginal of S_{obs} is not matched well in the simulated S values, this suggests model mis-specification

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard
- Regularization methods cheaper, but less studied and so less well understood
- Interpretable output
- Useful for model criticism: if some marginal of S_{obs} is not matched well in the simulated S values, this suggests model mis-specification

- Assumes there *is* a useful low dimensional summary

Discussion

Some observations:

- The two-stage method of Nunes and Balding seems to be a gold standard
- Regularization methods cheaper, but less studied and so less well understood
- Interpretable output
- Useful for model criticism: if some marginal of S_{obs} is not matched well in the simulated S values, this suggests model mis-specification

- Assumes there *is* a useful low dimensional summary
- Cost and scalability can be an issue

Projection methods

Projection methods:

- Start with a vector of data features Z, and try to find an informative lower dimensional projection

Projection methods

Projection methods:

- Start with a vector of data features Z, and try to find an informative lower dimensional projection
- To find projection, training data $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n_{\text{train}}$ are created from prior and model, and a projection is identified

Projection methods

Projection methods:

- Start with a vector of data features Z, and try to find an informative lower dimensional projection
- To find projection, training data $(\theta_i, \mathcal{D}_i), i = 1, \ldots, n_{\text{train}}$ are created from prior and model, and a projection is identified

1. Wegmann et al. (2009) Partial least squares

PLS aims to find linear combinations of covariates that have high covariance with responses and are uncorrelated with each other. Covariates are Z, and responses are $\theta = (\theta^{(1)}, \ldots, \theta^{(p)})$.

The ith PLS component $u_i = \alpha_i^T Z$ maximises

$$\sum_{j=1}^{p} \mathrm{Cov}(u_i, \theta^{(j)})^2$$

subject to $\mathrm{Cov}(u_i, u_j) = 0, j < i$, and a normalization such as $\alpha_i^T \alpha_i = 1$.

The ith PLS component $u_i = \alpha_i^T Z$ maximises

$$\sum_{j=1}^{p} \mathrm{Cov}(u_i, \theta^{(j)})^2$$

subject to $\mathrm{Cov}(u_i, u_j) = 0, j < i$, and a normalization such as $\alpha_i^T \alpha_i = 1$.

There are several methods for this (e.g. *pls* in R); they can give different results due to different normalizations.

2. Fearnhead and Prangle (2012) Linear regression

Fit a linear model to the training data: $\theta \sim \mathrm{N}(AZ + b, \Sigma)$, and use the resulting vector of parameter estimates $\hat{\theta} = AZ + b$ is used as summary statistics.

2. Fearnhead and Prangle (2012) Linear regression

Fit a linear model to the training data: $\theta \sim \mathrm{N}(AZ + b, \Sigma)$, and use the resulting vector of parameter estimates $\hat{\theta} = AZ + b$ is used as summary statistics.

Remark: this paper is wonderful . . . many other approaches, and it is an RSS discussion paper.

2. Fearnhead and Prangle (2012) Linear regression

Fit a linear model to the training data: $\theta \sim \mathrm{N}(AZ + b, \Sigma)$, and use the resulting vector of parameter estimates $\hat{\theta} = AZ + b$ is used as summary statistics.

Remark: this paper is wonderful . . . many other approaches, and it is an RSS discussion paper.

The package *abctools* in R implements these methods. See Prangle et al. (*Aust N Z J Stat*, 2014) for further details