Simulate Data Walkthrough

2025-09-19

```
set.seed(123)
```

Simulating Data

The functions required to simulate an example data set can be found in the file 'Simulate Data.R', which we load in below.

```
source('Simulate Data.R')
```

This file loads in example hyper-parameters (stored in list hyperParams) and survey details (stored in list dataParams) in order to simulate an example data set. To check that these have loaded in correctly, the following code should run without error:

```
data <- simulateData(dataParams, hyperParams)
head(data$dataParams$data_real) # View example data
```

```
##   Site Time Sample Plate       Ct
## 1    1    1      1     1 39.82511
## 2    1    1      1     1       NA
## 3    1    1      1     1 39.80312
## 4    1    1      1     1       NA
## 5    1    1      1     1 39.46689
## 6    1    1      2     1 38.62479
```

```
tail(data$dataParams$data_real)
```

```
##      Site Time Sample Plate       Ct
## 4995   10   20    999   200 37.98850
## 4996   10   20   1000   200 36.01068
## 4997   10   20   1000   200       NA
## 4998   10   20   1000   200 36.16541
## 4999   10   20   1000   200 36.58394
## 5000   10   20   1000   200 35.52392
```

In the following, we explain how to set up the lists dataParams and hyperParams.

dataParams

This list contains information regarding the survey design, with the following parameters:

1. ncovb: (integer) the number of site level covariates to generate

2. ncovw: (integer) the number of sample level covariates to generate
3. n: (integer) the number of sites in the study
4. nT: (integer) the number of time-points in the study
5. M: (integer vector) i-th entry denotes the number of samples taken at site 1+((i-1) %/% nT) and time-point 1+((i-1) %% nT) (for %/% and %% the quotient and remainder operators).
6. K: (integer vector) i-th entry denotes the number of replicates used of sample i
7. w_standards: (numeric vector) the concentrations of DNA used in the standards per plate
8. K_standards: (integer) the number of replicates used per standard concentration
9. CT.max: (double) the maximum number of cycles per PCR run

These values are then added as named elements to the list dataParams.

```
{
  ncovb = 2 # num site level covariates
  ncovw = 2 # num sample level covariates
  n = 10 # num sites
  nT = 20 # num time-points
  M = rep(5, n*nT) # num samples per site
  K = rep(5, sum(M)) # num replicates per sample
  w_standards = c(3e+07, 3e+06, 3e+05, 3e+04, 3e+03, 3e+02, 3e+01)
  K_standards = 3 # num replicates per standard concentration
  CT.max = 40 # Censoring limit
}

dataParams <- list(ncovb = ncovb,
                   ncovw = ncovw,
                   n = n,
                   nT = nT,
                   M = M,
                   K = K,
                   w_standards = w_standards,
                   K_standards = K_standards,
                   CT.max = CT.max
)
```

hyperParams

This list contains details about the parameters controlling qPCR outputs and the factors affecting DNA concentrations through space and time. These parameters are explained in full detail in the Manuscript in the repo, we give only brief descriptions here. The parameters in this list include:

1. tau: (double > 0) standard deviation for the noise associated with the time series (τ in the paper)
2. sigma: (double > 0) standard deviation for the noise associated with sampling (σ in the paper)
3. tau2.1: (double > 0) variance for the the distribution of DNA across sites at the first time-point (τ_1^2 in the paper)
4. betab: (double, vector of length ncovb) the site covariate coefficients (β_b in the paper)
5. betaw: (double, vector of length ncovw) the sample covariate coefficients (β_w in the paper)
6. betab0: (double) the mean log-DNA concentration across sites at the first time-point ($\beta_b, 0$ in the paper)
7. alpha1.0: (double) the mean (across plates) for the intercept coefficient in the plate-regression (α_0^1 in the paper)
8. alpha2.0: (double) the mean (across plates) for the slope coefficient in the plate-regression (α_0^2 in the paper)

9. sigma_alpha1: (double > 0) the standard deviation (across plates) for the intercept coefficient in the plate-regression (σ_α in the paper)
10. sigma_alpha2: (double > 0) the standard deviation (across plates) for the slope coefficient in the plate-regression (σ_α in the paper)
11. rho: (double, vector) i-th entry denotes the time series coefficient for site i (ρ in the paper)
12. a: (double) the intercept for the log variance associated with CT heteroscedasticity (a_1 in the paper)
13. b: (double) the slope for the log variance associated with CT heteroscedasticity (a_2 in the paper)
14. lambda0: (double) determines the mean of contaminating DNA concentrations
15. sd_lambda: (double > 0) the standard deviation of contaminating DNA concentrations
16. p0: (double $\in (0, 1)$, vector of length 2) first entry denotes probability that a replicate is contaminated, second entry denotes probability that a replicate is inhibited ((p_c, p_h) in the paper)
17. multiplier: (double $\in (-1, 0)$) the effect of inhibition, reduced the amount of DNA in a replicate by this proportion

These values are then added as named elements to the list hyperParams.

```
hyperParams <- list(tau = 1, # sd time series
                    sigma = 1, # sd samples
                    tau2.1 = 1, # variance across sites when t=1
                    betab = c(1, -1), # site covariate coeffs
                    betaw = c(1, -1), # sample covariate coeffs
                    betab0 = 6, # intercept across sites when t=1
                    alpha1.0 = 44, # mean plate intercept
                    alpha2.0 = -1.7, # mean plate slope
                    sigma_alpha1 = .1, # sd plate intercept
                    sigma_alpha2 = .01, # sd plate slope
                    rho = rep(1, n), # time series coeff
                    a = 0.2, # plate variance intercept
                    b = -0.25, # plate variance slope
                    lambda0 = 3e+3, # mean contamination concentration
                    sd_lambda = 100, # sd contamination concentration
                    p0 = c(0.05, 0.1), # prob contamination and inhibition
                    multiplier = -9/10 # inhibition effect
)
```

Modelling assumptions

We make some assumptions in the simulated data sets (that are not necessary for real data and can be modified/removed in the original code if required):

1. Half the covariates generated for the sites and samples will be continuous and half will be binary.
2. With respect to the standards, we assume that the same number of replicates are being used for each concentration.
3. Each sampling occassion is analysed on a single plate, so that the study will comprise n x nT plates in total.

We also note that the effect of inhibition is proportional to the amount of DNA in the sample. A multiplier of -0.5 has the effect of removing 50% of the DNA in the replicate.

qPCR data output

simulateData outputs a list of two named elements:

1. trueParams (list)
2. dataParams (list)

The information contained in trueParams is 'latent' meaning that in a real data set these would be unknown (for example the true concentrations of DNA in the environment is unknown, but we record these to compare our estimates to). The information in dataParams is observed and what we expect to record during a true eDNA qPCR survey.

The outputs in trueParams include:

1. l_true: (vector) a vector of the true log-DNA concentrations. The i-th entry corresponds to site 1+((i-1) %/% nT) and time-point 1+((i-1) %% nT)
2. v_true: (vector) a vector of the true log-DNA concentrations in each sample. The i-th entry corresponds to the i-th sample as labelled in `data$dataParams$data_real`
3. rho_true: (double, vector of length n) rho from hyperParams
4. tau_true: (double) tau from hyperParams
5. tau1_true: (double) the square root of tau2.1 from hyperParams
6. alpha1_true: (double, vector of length n x nT) the i-th value corresponds to the intercept in the plate regression of the i-th plate
7. alpha2_true: (double, vector of length n x nT) the i-th value corresponds to the slope in the plate regression of the i-th plate
8. a_true: (double) a from hyperParams
9. b_true: (double) b from hyperParams
10. beta_b_true: (double, vector of length ncovb) betab from hyperParams
11. beta_w_true: (double, vector of length ncovw) betaw from hyperParams
12. betab0_true: (double) betab0 from hyperParams
13. delta_true: (binary, vector of length the number of rows in `data$dataParams$data_real`) returns 1 if replicate amplified (0 otherwise) for replicates from collected samples. The i-th value corresponds to the i-th row of `data$dataParams$data_real`
14. delta_star_true: (binary, vector of length the number of rows in `data$dataParams$data_standrd`) returns 1 if replicate amplified (0 otherwise) for replicates from standards. The i-th value corresponds to the i-th row of `data$dataParams$data_standard`
15. sigma_true: (double) sigma from hyperParams
16. lambda_true: (double, vector of length the number of rows in `data$dataParams$data_real`) i-th value is the quantity of DNA concentration added to (value is positive in case of contamination) or removed from (value is negative in case of inhibition) the replicate in the i-th row of `data$dataParams$data_real`. Value is 0 if replicate is neither contaminated or inhibited
17. lambda_star_true: (double, vector of length the number of rows in `data$dataParams$data_standard`) i-th value is the quantity of DNA concentration added to (value is positive in case of contamination) or removed from (value is negative in case of inhibition) the replicate in the i-th row of `data$dataParams$data_standard`. Value is 0 if replicate is neither contaminated or inhibited
18. gamma_true: (binary, vector of length the number of rows in `data$dataParams$data_real`) returns a 1 if lambda_true is not zero (0 otherwise)
19. gamma_star_true: (binary, vector of length the number of rows in `data$dataParams$data_standard`) returns a 1 if lambda_star_true is not zero (0 otherwise)
20. p0_true: (double $\in (0, 1)$, vector of length 2) p0 from hyperParams
21. lambda0_true: (double) lambda from hyperParams
22. sd_lambda_true: (double > 0) sd_lambda from hyperParams

The outputs in dataParams include:

1. data_real: (data frame) a data frame with columns Site (integer), Time (integer), Sample (integer), Plate (integer), and Ct (double > 0). Each row corresponds to a replicate from a sample collected in the environment. If a replicate fails to amplify, its Ct entry is NA. Each sample value corresponds to a unique site and time-point

2. data_standard: (data frame) a data frame with columns Plate (integer), Quantity (double > 0), and Ct (double > 0). Each row corresponds to a replicate from the standards. If a replicate fails to amplify, its Ct entry is NA. The Plate values here correspond to the same plates as in data_real. The Quantity denotes the DNA concentration in the standard replicate.
3. X_b: (array, dimensions (n x nT x ncovb)) an array of the covariate observations at the site level. The i,j-th row corresponds to site i and time-point j
4. X_w: (matrix, dimensions (num_samples x ncovw)) a matrix of the covariate observations at the sample level. The i-th row corresponds to sample i from data_real
5. ncovb: (integer) ncovb from dataParams
6. ncovw: (integer) ncovw from dataParams
7. id_site: (integer, vector of length num_samples) i-th entry denotes the sampling occasion j that sample i was collected from. Sampling oaccasion j corresponds to site 1+((j-1) %/% nT) and time-point 1+((j-1) %% nT)
8. id_sample: (integer, vector of length num_replicates) i-th entry denotes the sample that the i-th row of data_real is associated with
9. id_site_l: (integer, vector of length n x nT) i-th entry denotes the site associated with sampling occasion i (1+((i-1) %/% nT))
10. id_site_time: (integer, vector of length n x nT) i-th entry denotes the time associated with sampling occasion i (1+((i-1) %% nT))
11. P: (integer, vector of length num_replicates) i-th entry corresponds to the plate that the replicate from the i-th row of data_real was analysed on
12. P_star: (integer, vector of length num_replicates_star) i-th entry corresponds to the plate that the replicate from the i-th row of data_standard was analysed on
13. numP: (integer) total number of plates used during study (equal to maximum value in data_real$Plate)
14. num_sites: (integer) n from dataParams
15. nT: (integer) nT from dataParams
16. num_samples: (integer) total number of samples used during study, excluding standards (equal to maximum value in data_real$Sample)
17. num_replicates: (integer) total number of replicates in study, excluding standards (equal to number of rows in data_real)
18. num_replicates_star: (integer) total number of replicate standards in study (equal to number of rows in data_standard)
19. CT.max: (double) CT.max from dataParams

Note:

The code in 'Model Codes.R' is not currently set up to handle missing covariate values.