Abstract

Here we represent repdoducible code for the simulation study appearing in the paper 'Spatio-temporal bivariate statistical models for atmospheric trace-gas inversion', Section 4.1. The code requires the installation of two in-house developed packages for this application, *hmc* and *atminv*. The vignette itself is not 'polished', but gives the basic requirements for reproducing the figures and values given in the main text.

1 Setup

To run the simulation study, you need to first install the packages 'atminv' and 'hmc'. You can do this as follows:

```
library(devtools)
install_github("andrewzm/hmc")
install_github("andrewzm/atminv")
```

This only needs to be done once. Now that we have the development packages installed, we can now load the others that we will need. The first two, `ggplot2` and `grid` are plotting purposes. The package `dplyr` is used for fast table manipulation and for 'piping' a sequence of commands. The package `Matrix` is needed for taking advantage of sparsity in some operations and `tidyr` is needed for rearranging tables. The package `gstat` is needed for variogram modelling of the flux field. Finally the in-house developed packages `hmc` and `atminv` are used for implementing the Hamiltonian Monte Carlo sampler and the EM algorithm for parameter estimation, respectively.

```
library(ggplot2)
library(grid)
library(dplyr)
library(Matrix)
library(tidyr)
library(gstat)
library(hmc)
library(atminv)
# load_all("../../../pkgs//hmc")
#library(devtools)
#load_all("..")
#library(scales) # for format_format
#load_all("../../../../../CurrentProjects/PostDoc Bristol/R Code/pkg/MVST")
```

We will now set up our simulation. Here we will only be concerned with the model 'full' ('full_big' analyses the case with 1000 observations and 'diag' the case where the flux field is uncorrelated), which can be misspecified (`misspecification = 1`) or not (`misspecification = 0`). Recall the by misspecification here we imply that the flux field is indeed spatially correlated, but that we will model is as being uncorrelated. Below we set up the spatio-temporal grid and establish the parameters of the observation process and the mole-fraction discrepancy spaio-temporal field:

```
###------------------
### Parameters
###------------------
model = "full" ## Either sparse or full or full_big or diag
```

```
misspecification = 0
#set.seed(25) # 25 , T = 400 or 15/200
ds <- 0.2             # 0.2 spacing.
                      # NB: If we change this we need to
                      # change the round() command further down
smin = -10 + ds/2     # first gridcell centre
smax =  10 - ds/2     # last gridcell centre

s_axis <- round(seq(smin,smax,by=ds),1) # create s-axis
if(model %in% c("full","diag")) {
  t_axis <- 1:100                       # create t-axis
  m_obs <- 6                            # number of obs. (including val.)
} else {
  t_axis <- 1:100                       # create t-axis
  m_obs <- 1000                         # number of obs.
}

ns <- length(s_axis)                    # no. of gridcells
nt <- length(t_axis)                    # no. of time points
st_grid <- expand.grid(s=s_axis,        # ST-grid (long format)
                       t=t_axis) %>%
            data.frame()

sigma_eps_true <- 10                    # observation error std.
if(model %in% c("full","full_big")) {
  sigma_zeta_true <- 50                 # discrepancy marginal std.
} else {
  sigma_zeta_true <- 10                 # discrepancy marginal std.
}

theta_t_true <- 0.8     # temporal correlation parameter ('a' in text)
theta_s_true <- 1       # spatial range parameter ('d' in text)
```

Now we are ready to create a stochastic process, the realisations of which exhibit similar characteristics to what we will be studying in the real example. Recall that the stochastic process we use is:

$$b_t(s,u \mid v_t(s)) \equiv \exp\left(-\frac{(u-s)^2}{2v_t(s)^2}\right) I\big(|u-s| < |v_t(s)|\big) J(s,u), \tag{1}$$

where

$$J(s,u) \equiv \begin{cases} I[(u-s) \geq 0]; & v_t(s) \geq 0, \\ I[(u-s) \leq 0]; & v_t(s) < 0, \end{cases}$$

where $v_t(s)$ from a Gaussian process with separable spatio-temporal covariance structure and $I(\cdot)$ is the indicator function. In (1), the exponential function describes a bell-shaped curve centred at $u = s$, while the indicator function truncates this curve at $u = s \pm v_t(s)$. The third term, $J(s,u)$, then truncates the bottom half of the function if $v_t(s) \geq 0$ and the upper half otherwise. The function $(s,u \mid v_t(s))$ is implemented as follows

```
###-----------------
### Transition kernel
###-----------------
```

```
# p is a ST process and reflects the std of the truncated Gaussian.
# This problem ONLY works if b is of relatively local scope. Once we
# have b which has a very large scope we get oscillations/instability

b <- function(s,u,p) {
  absp <- max(abs(p),0.2)
  absp*sqrt(2*pi) * dnorm(u,mean = s, sd =absp) *
    ((sign(p) == sign(u-s)) | (u-s) == 0) *
    (abs(u - s) < absp)
}
```

while the spatio-temporal Gaussian parameter is simulated from a separable field as follows:

```
## Sample the "wind" vector
Q_s <- GMRF_RW(n = length(s_axis),
               order = 2,
               precinc = 2000)@Q +
          0.001*.symDiagonal(length(s_axis))  # spatial precision
Q_t <- GMRF_RW(n = length(t_axis),
               order = 1,
               precinc = 20)@Q +
          0.1*.symDiagonal(length(t_axis))    # temporal precision

Q_full = as(kronecker(Q_t,Q_s),"dgCMatrix")   # spatio-temporal precision

G <- GMRF(mu = matrix(rep(0,nrow(Q_full))),   # GMRF with final precision
          Q = Q_full,n=nrow(Q_full))

# Load the seed we used to simulate this parameter
data(sim.Random.seed)
#load("~/Desktop/Chemometrics_results/sim.Random.seed.rda")

# Now simulate this parameter by sampling from the GMRF
st_grid$p <- sample_GMRF(G,reps = 1)
```

If we want we can take a look at what the realisation of the parameter $v_t(s)$ looks like through

```
print(LinePlotTheme() +
        geom_tile(data=st_grid,aes(s,t,fill=p)) +
        scale_fill_gradient2(low="blue",high="red",mid="white")
      )
```

the result of which is depicted in Figure 1

2 Process and observation simulation

2.1 Simulating the flux field

Now that we have the parameters in place, we can simulate our dataset. We first re-set the seed to '1', then construct a semi-variogram with the parameters identical to those estimated from the Emissions data, before simulating our vector $\mathbf{Y}_f$:

```
print(LinePlotTheme() +
        geom_tile(data=st_grid,aes(s,t,fill=p)) +
        scale_fill_gradient2(low="blue",high="red",mid="white")
      )
```

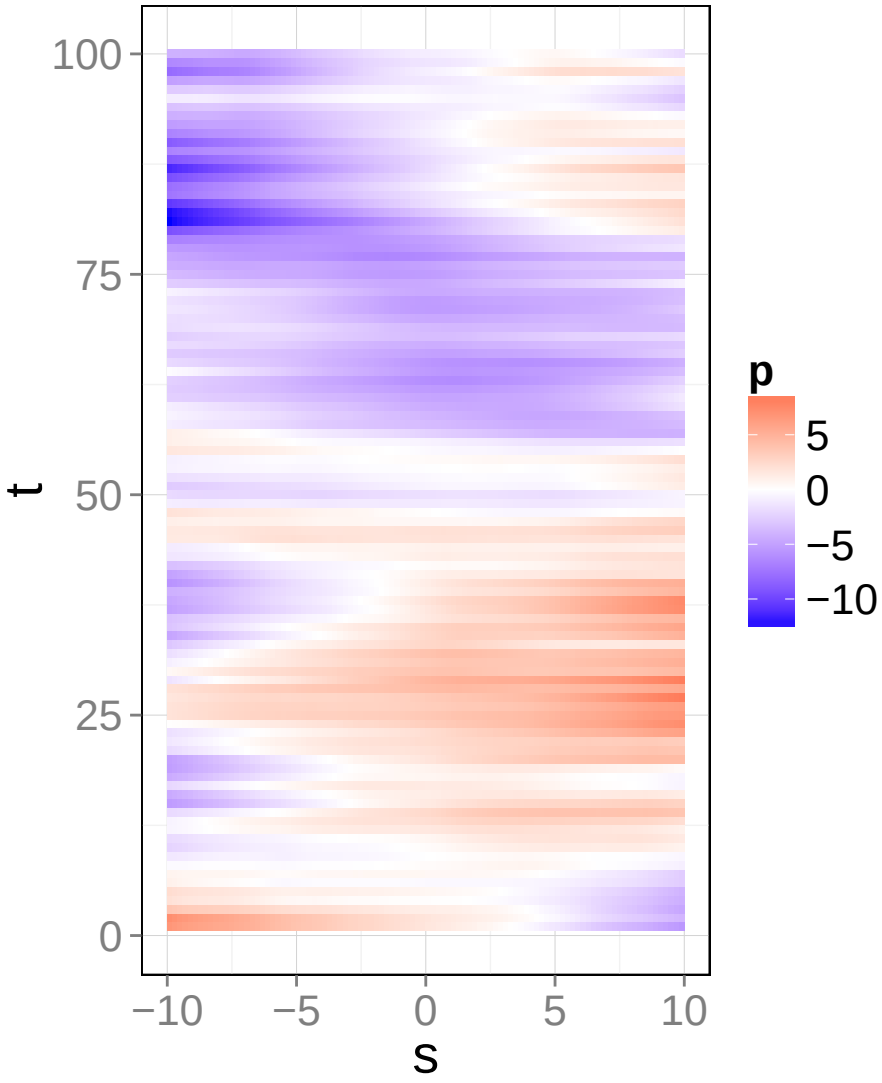

Figure 1: The parameter $v_t(s)$ simulated from the separable spatio-temporal process

```
###------------------
### Lognormal flux field
###------------------
set.seed(1)
variogram_model <- vgm(range = 3.334,            # Construct spherical semi-variogram
                       nugget = 0.00533,
                       psill = 0.80429,
                       model= "Sph")
S_f_log <- variogramLine(variogram_model,        # Find covariance matrix Sigma_f
                         dist_vector = sp::spDists(matrix(s_axis),
                                                   matrix(s_axis)),
                         covariance = TRUE)
mu_f_log <- matrix(rep(5,length(s_axis)))        # Construct mu_f
Yf_sim <- exp(mu_f_log + t(chol(S_f_log)) %*%    # Simulate Y_f
                rnorm(n = length(s_axis)))
st_grid <- st_grid %>%                           # Append Y_f to data frame
          left_join(data.frame(s=s_axis,
                               Yf = Yf_sim))

## Joining by:  "s"

if(misspecification)  {                           # If we are assuming misspecification
    S_f_log <- diag(diag(S_f_log))                # We over-write Sigma_f to be diagonal
}
```

2.2 Simulating the mole fraction field

Since the spatio-temporal discrepancy term can be hard to simulate from withour further approximations, we will just simulate it at the observation locations. Therefore, the mole fraction (excluding the discrepancy) is just a linear transformation of the flux field. For each space-time location we find the SRR between the whole domain (the source) and that point (the receptor) and multiply it by the the flux in the that gridcell:

```
###------------------
### Mole fraction field
###------------------
## Since we cannot put the discrepancy everywhere (too large),
## we will just put it at the observation locations
st_grid <- st_grid %>%
  group_by(t,s,p) %>% # for each space-time location
  summarise(Yf = Yf,  # find the mole-fraction by finding the SRR
            Ym = sum(b(s =s, u = s_axis,p = p) *
                       Yf_sim * ds))
```

2.3 Simulating the observations

We randomly choose `m_obs` observations from the spatial grid (excluding the lower and upper 10 grid cells) and assume these are observed. We replace the 6-th observation location with $s = 0.3$ that will be used for validation.

```
###------------------
### Observations
###------------------
if(model %in% c("full","diag")) {
      s_obs <- data.frame(s = sample(s_axis[-c(1:10,(ns-10):ns)],
                                     size = m_obs,
                                     replace=F),
                          m = 1:m_obs  )
    new_obs <- 0.3
    s_obs[6,]$s <- new_obs
} else {
  s_obs <- data.frame(s = sample(s_axis,size = m_obs, replace=T),
                      m = 1:m_obs)
}
```

Now we merge the observation data frame with the spaio-temporal grid and add the observation error, before sorting the data frame by time and space:

```
s_obs <- s_obs %>%
            left_join(st_grid) %>%
            mutate(z = Ym + rnorm(n = length(Ym),
                                  sd = sigma_eps_true)) %>%
            arrange(t,s)

## Joining by:  "s"

Qobs <- sigma_eps_true^(-2) * .symDiagonal(nrow(s_obs))
```

To add the discrepancy, we first compuate the spatio-temporal covariance matrix at the observation (space-time) locations using the `corr_zeta_fn` function in the `atminv` package, find its Cholesky decomposition and then use it to simulate the discrepancy at the required locations. For when we have 2000 observations (`model = 'full_big'`) we find the covariance matrix at every space-time grid location and use assign the discrepancy to the observations at the grid level. The variable `C_m` is a matrix that maps the location of the mole fraction observations to the mole fraction prediction grid. Note that for the 'full' model this is just the $m \times m$ since with this model we are choosing not predicting the mole fraction in every grid cell:

```
## Now add the discrepancy
if(model %in% c("full","diag")) {

    corr_zeta_true <- corr_zeta_fn(s_obs$s[1:m_obs],
                                   t_axis,
                                   theta_t_true,
                                   theta_s_true)
    S_zeta_true <- sigma_zeta_true^2 * corr_zeta_true
    chol_S_zeta_true <- chol(S_zeta_true)

    s_obs <- s_obs %>%
      mutate(dis = t(chol_S_zeta_true) %*%  rnorm(n = nrow(s_obs)),
             z = z + dis)

    C_m <- .symDiagonal(nrow(s_obs))
```

```
} else if(model == "full_big") {

    corr_s_mat <- function(theta_s) corr_s(s = s_axis,theta_s = theta_s)
    corr_t_mat <- function(theta_t) corr_t(t = t_axis,theta_t = theta_t)
    d_corr_s_mat <- function(theta_s) d_corr_s(s = s_axis,theta_s = theta_s)
    d_corr_t_mat <- function(theta_t) d_corr_t(t = t_axis,theta_t = theta_t)

    C_idx <- st_grid %>%
      as.data.frame() %>%
      select(s,t) %>%
      mutate(n = 1:nrow(st_grid)) %>%
      left_join(s_obs,.)

    C_m <- sparseMatrix(i=1:nrow(C_idx),
                        j = C_idx$n,
                        x=1,
                        dims=c(nrow(s_obs),nrow(st_grid)))

    chol_S_zeta_true <-  sigma_zeta_true *
                             kronecker(chol(corr_t_mat(theta_t_true)),
                                       chol(corr_s_mat(theta_s_true)))

    s_obs <- s_obs %>%
        mutate(dis = as.vector(C_m %*%
                                 (t(chol_S_zeta_true) %*%
                                    rnorm(n = ns*nt))),
               z = z + dis)
}
```

2.4 Illustrative plots

Here we provide the code for generating the two plots in the paper (Figure 3). The first, in Figure 2, shows the average mole fraction and the flux field superimposed:

```
if(model == "full") {
    X <- group_by(st_grid,s) %>%
        summarise(Yf = Yf[1],Ym_av = mean(Ym)) %>%
        gather(process,value,-s)
    g <- LinePlotTheme() +
          geom_line(data=subset(X,!(s==new_obs)),
                    aes(x=s,y=value,linetype=as.factor(process)))+
          geom_segment(data=s_obs[1:5,],
                       aes(x=s, xend=s, y = 0, yend = 50),
                        arrow=arrow(length=unit(0.1,"cm"))) +
        ylab("") +
        scale_linetype_discrete(guide=guide_legend(title="process"),
                                labels=c("Yf (g/s/degree)","Ym (ppb)")) +
        xlab("s (degrees)")
    ggsave(g,filename = "../../Sim_plot.png",width=10,height=4)
}
```

```
print(g)
```

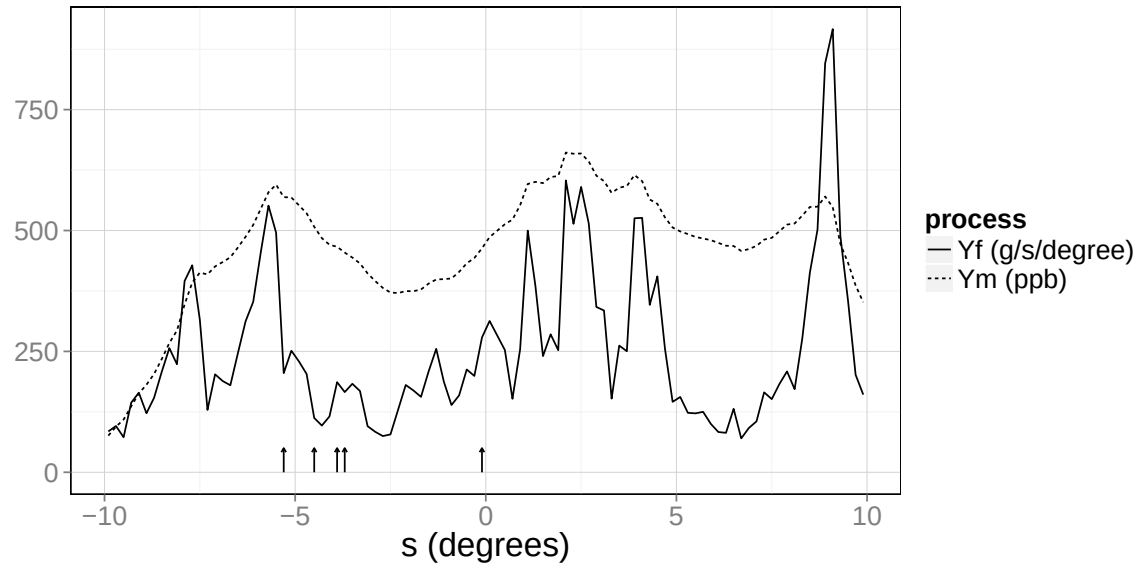

Figure 2: A sample realisation of the flux field (solid line), the resulting time-averaged mole-fraction field (dashed line) and the five observation locations (arrows).

The second, in Figure 3, shows the SRR at each observation location:

```
if(model %in% c("full","diag")) {
  df_for_B <- s_obs
} else {
  df_for_B <- st_grid
}

# TRUE B
B_true <- plyr::ddply(df_for_B,c("t","s"),function(df) {
  b = b(s=df$s[1],u=s_axis,p=df$p[1])}) %>%
  select(-s,-t) %>%
  as.matrix()*ds

B <- B_true
if(model == "sparse") {
  B <- as(B,"dgCMatrix")
}

B_true_df <- plyr::ddply(df_for_B,c("t","s"),function(df) {
  b = b(s=df$s[1],u=s_axis,p=df$p[1])}) %>%
  gather(s_grid,b,-t,-s) %>%
  separate(s_grid, into = c("V","s_fine"),sep="V") %>%
  select(-V) %>%
  mutate(s_fine = round(s_axis[as.numeric(s_fine)],1))# %>%
#left_join(model_pred_em,by = c("t","s","s_fine"))
```

```
if(model == "full") {
  ## Plot the source-receptor relationship at each observation

  ## The following code uses colours and shows the SRR on one plot
  #   B_plot <- LinePlotTheme() +
  #             geom_tile(data=subset(B_true_df ,b>0),
  #                       aes(x=s_fine,y=t,fill=as.factor(s),alpha=b)) +
  #     scale_alpha_continuous(range=c(0,1)) + xlab("u") +
  #     scale_fill_discrete(guide=guide_legend(title="s")) +
  #     coord_fixed(xlim=c(-10,10),ratio = 0.2)

  ## The following code shows one SRR per plot
  B_plot <- LinePlotTheme() +
    geom_tile(data=subset(B_true_df,b>0 & !(s==new_obs)),
              aes(x=s_fine,y=t,alpha=b),fill="black") +
    scale_alpha_continuous(guide=guide_legend(title="s/ng")) +
    scale_y_reverse()+
    scale_fill_discrete(guide=guide_legend(title="s")) +
    coord_fixed(xlim=c(-10,10),ratio = 0.5)  +
    facet_grid(~s) +
    theme(panel.margin = unit(1.5, "lines")) +
    xlab("u (degrees)") +
    ylab("t (2 h steps)")
  ggsave(filename = "../../B_plot.png",width=12)
}
```

3 EM algorithm

In this section we show the code for carrying inference on the bivariate field $(\mathbf{Y}_f, \mathbf{Y}_m)'$. We first compute the mean flux density, which we will then also use as the initial value for the conditional expectation of $\mathbf{Y}_f$ in the gradient descent:

```
mu_f <- matrix(exp(mu_f_log + 0.5*diag(S_f_log)))
```

We then set the settings for the Laplace approximation. These vary accoring to the model being used. For the model 'full', we alter the matrices so that the validation point is excluded. :

```
###-----------------------------------------------------
### Laplace method -- use with caution because of mode close to zero
###-----------------------------------------------------
n_EM <- 100

if(model == "full") {
  s_mol = s_obs$s[1:m_obs]          # prediction locs for mol fraction
  Y_init = c(mu_f,s_obs$z)          # initial expectation value for Y
  theta_init = c(1000,0.2,0.2)      # initial parameter vector
                                    # where theta = [sigma_zeta^2, theta_t, theat_s]
```

```
print(B_plot)
```

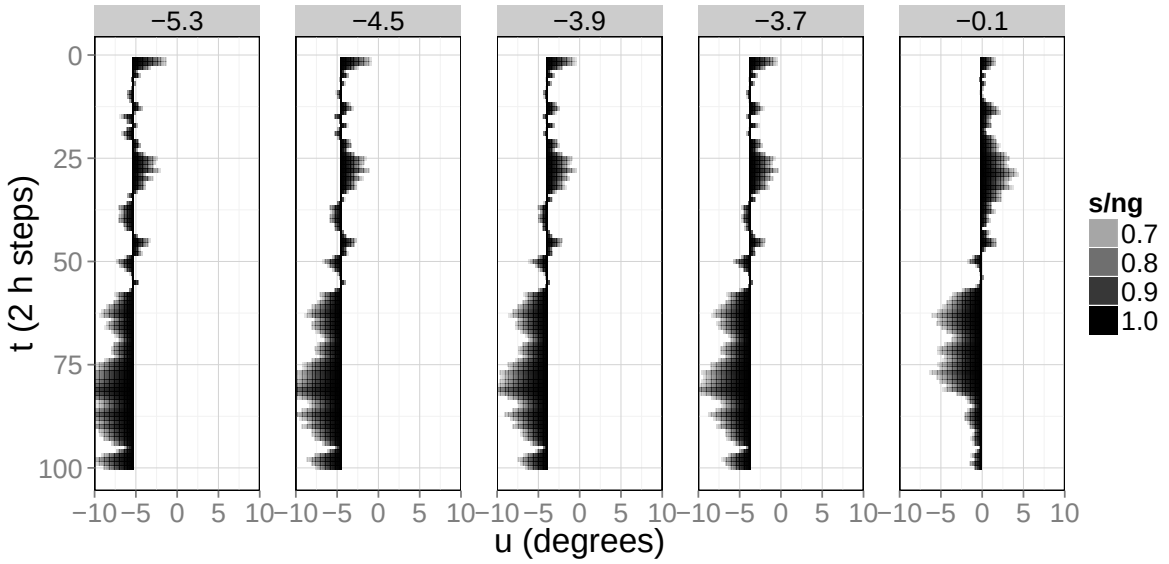

Figure 3: The source-receptor relationship $b_t(s,u)$ synthesised at five observation locations $s \in D_m^O = \{-5.3°, -4.5°, -3.9°, -3.7°, -0.1°\}$. Note that $b_t(s,u) = 0$ for $u > 4.3$

```
  # Keep station out for validation
  rm_idx <- seq(6,nrow(C_m),by=6) # index to be removed
  s_obs_old <- s_obs              # save old observation location data frame
  s_obs <- s_obs[-rm_idx,]        # new observation data frame
  C_m <- C_m[-rm_idx,]            # new incidence matrix
  Qobs <- Qobs[-rm_idx,-rm_idx]   # new (observation) precision matrix

} else if (model == "full_big") {
    s_mol = s_axis                # prediction locs for mol fraction
    Y_init = c(mu_f,st_grid$Ym)   # initial expectation value for Y
    theta_init = c(1000,0.2,0.2)  # initial parameter vector
                                  # where theta = [sigma_zeta^2, theta_t, theat_s]
}
```

We are now ready to call the main function of the `EM_alg` package, which returns a function that can be iterated. The inputs to this function are given in-line:

```
EM_alg <- EM(s_obs = s_obs,                 # observation data frame
        C_m = C_m,                          # incidence matrix
        Qobs = Qobs,                        # observation precision matrix
        B = B,                              # SRR matrix
        t_mol = t_axis,                     # time prediction locs
        s_mol = matrix(s_mol),              # spatial prediction locs
        S_f_log = S_f_log,                  # cov. matrix of log(Yf)
```

```
        mu_f_log = mu_f_log,                  # expectation of log(Yf)
        Yf_thresh = 1e-4,                     # drop nodes with Yf < 1e-4
        Y_init =  Y_init,                     # initialise Yf
        theta_init = theta_init,              # initialise theta
        ind = which(!(colSums(B) == 0)),      # only consider indices with SRR>0
        n_EM = n_EM,                          # number of EM iterations
        model = model)                        # model we are using
```

To iterate the E- and M-steps we simply put the returned function in a loop, and indicate the maximum numer of gradient descents to carry out in the E- and M- steps. In this case we are going to limit the number of M-steps to 50. We also allow for a variable `fine_tune_E` that indicates on which iteration to carry out a second gradient descent at a higher convergence tolerance. This is particularly useful when the Hessian computed at 'convergence' is not positive definite due to the tolerance used.

```
for(i in 1:(n_EM-2)) {
  X <- EM_alg(max_E_it = 1e6,
              max_M_it = 50,
              fine_tune_E = (i==0))
}
```

4 HMC sampler

In this section we show the code for the HMC sampler, that fixes the parameters found using the EM algorithm above. We first compute the covariance matrix:

```
if(model == "full") {
  corr_zeta <- corr_zeta_fn(c(s_obs$s[1:(m_obs-1)],new_obs),t_axis,X$theta[2,n_EM],X$theta[3
  S_zeta <- X$theta[1,n_EM] * corr_zeta
  Q_zeta <- chol2inv(chol(S_zeta))
}
```

Now we formulate the equations required for the log-likelihood and the gradient of the log-likelihood of $\mathbf{Y}_f$. These are provided in the helper function `Yf_marg_approx_fns` and its arguments are explained in-line.

```
lap2 <- Yf_marg_approx_fns(s = s_obs,             # obs. locations
                           C_m = C_m,             # incidence matrix
                           Qobs = Qobs,           # obs. precision matrix
                           B = B,                 # SRR matrix
                           S_zeta = S_zeta,       # Sigma_zeta
                           mu_f_log = mu_f_log,   # expectation of log(Y_f)
                           S_f_log = S_f_log,     # covariance of log(Y_f)
                           ind=X$ind)             # only consider indices with SRR>0
```

Next, we set some parameters for the sampler. We set the number of steps per sample $L = 10$, the number of samples $N = 10000$, and the step-size $\Delta \in [0.066, 0.068]$ (below denoted as 'eps_gen'). Other variables are defined in-line:

```
M <- diag(1/(X$lap_approx$Yf^2)) # scaling matrix: puts variables on same scale

if(model == "full"){             # Function for generating Delta
  eps_gen <- function()
    runif(n=1,
          min = 0.066,
          max = 0.068)
} else if (model =="diag") {
  eps_gen <- function()
    runif(n=1,
          min = 0.0065,
          max = 0.0067)
}
L <- 10L                     # step-size
N <- 3                   # number of samples
q <- matrix(0,nrow(M),N)    # matrix for storing samples
qsamp <- X$lap_approx$Yf    # first sample
dither <- i <- count <- 1   # no dithering, initialise counts
```

The sampler is set up using the function `sampler` in the `hmc` package. We set a lower limit of 0; no samples less or equal to zero are allowed:

```
sampler <- hmc_sampler(U = lap2$logf,                  # log-likelihood
                       dUdq = lap2$gr_logf,            # gr. of log-likelihood
                       M = M,                          # scaling matrix
                       eps_gen = eps_gen,              # step-size generator
                       L = L,                          # number of steps
                       lower = rep(0,length(X$ind)))   # lower limits
```

We now run the HMC sampler by repeatedly iterating through it. We also can monitor the acceptance rate. For conciseness the output below is not produced. We then calculate the acceptance rate again after the HMC is run.

```
while(i < N) {
  qsamp <- sampler(q = qsamp)

  if(count  == dither) {
    q[,i] <- qsamp
    count = 1
    i <- i + 1
    print(paste0("Sample: ",i," Acceptance rate: ",(nrow(unique(t(q)))-1)/i))
  } else {
    count <- count + 1
  }
}
```

```
print(paste0("Sample: ",i," Acceptance rate: ",(nrow(unique(t(q)))-1)/i))

## [1] "Sample: 5 Acceptance rate: 0.2"
```

5 Results

In this section we show the code used to generate the results. First, we put the samples into a format we can work with (the variable 'Q2') and then produce a box pot of the samples of the flux field at each of the prediction locations:

```
print(g)
```