

Project1

Markov Chain Monte Carlo techniques

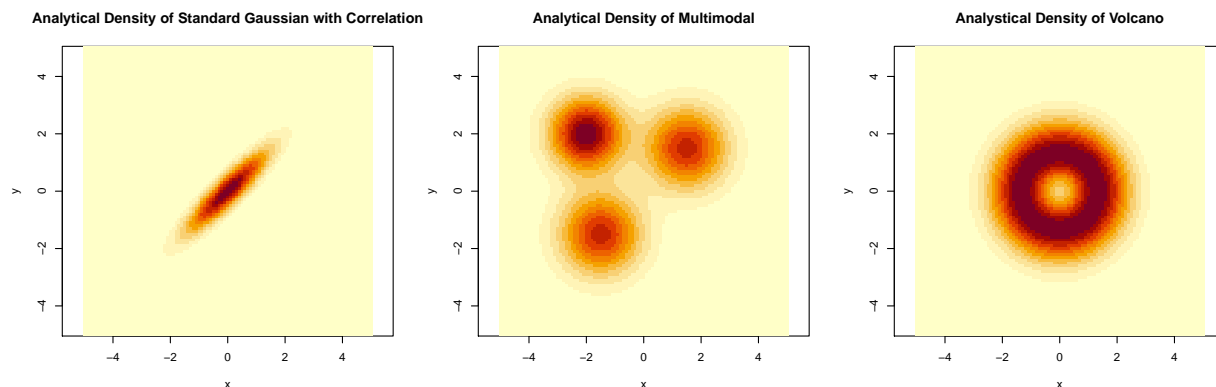
Florian Beiser, Yaolin Ge

1 Metropolis-Hastings for Bivariate Densities

We consider three different bivariate densities. 1. Standard Gaussian with correlation 2. Multimodal as mixture of three Gaussians 3. volcano (unnormalized)

1.1 Plotting

We start the project with the visualisation of the respective densities in $[-5, 5] \times [-5, 5]$. (Whyever R likes to put more white space left and right of the plots, but the middle areas are still representative.)



For the subsequent part, where different Metropolis Hastings MCMC algorithms will be implemented, we prepare with the skeleton that was discussed in the exercise class and uploaded. This defines the frame for all MCMC implementations under consideration which differ in the `proposal_func` and `acceptance_func`.

These implementation is not tweaked for optimal performance, e.g. log-scale could lead to better scaling or reduction of redundant calculations in the individual MCMC versions could lead to fast execution, but we want to the same skeleton for all algorithms to highlight similarities and differences. For the analysis of the outcomes, we use below's plotting functionality.

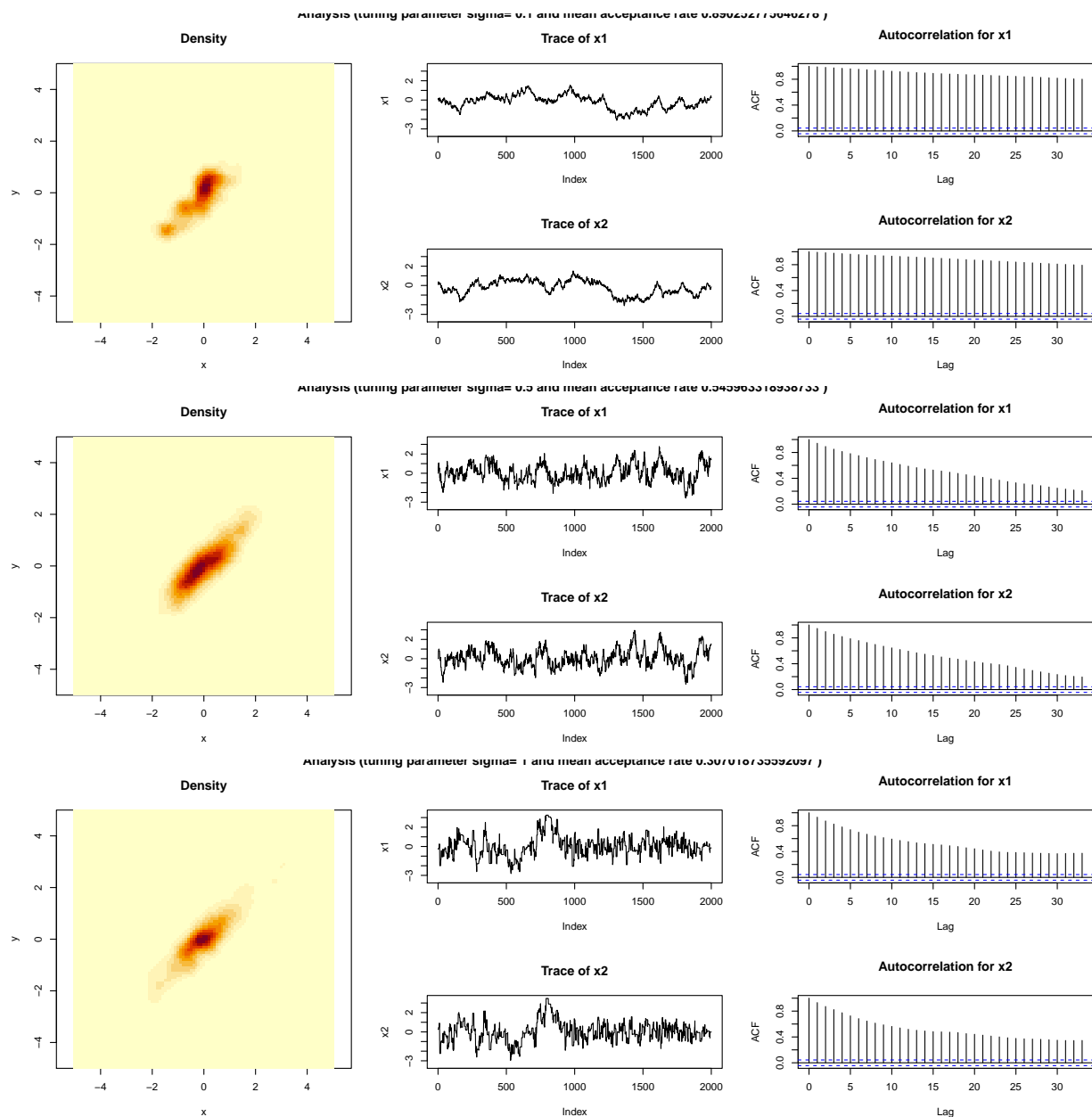
We will use the same input parameters for all subsequent MCMC calculations, the tuning parameters are of course not influenced.

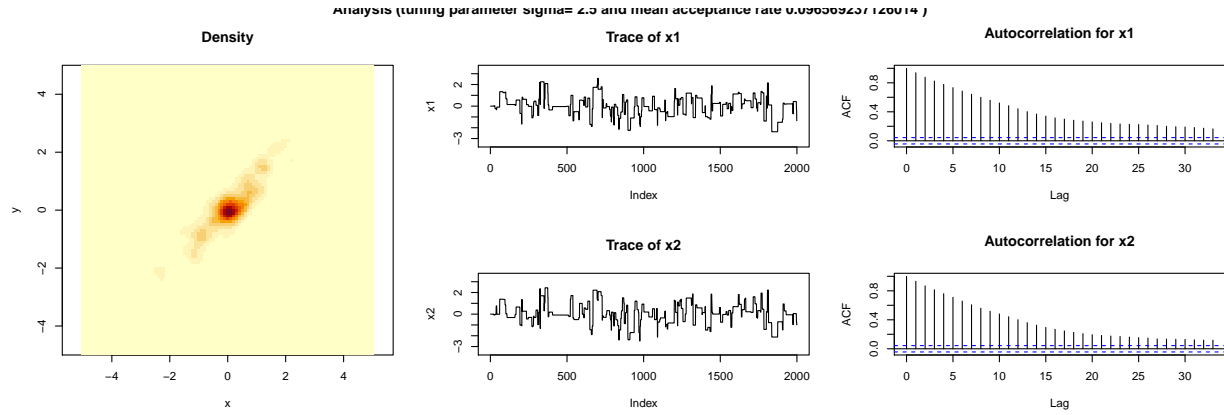
1.2 Random walk MH

The random walk MH uses a symmetric $\mathcal{N}(x_{i-1}, \sigma^2)$ distribution to generate the new proposal. Moreover, the general MH acceptance probability simplifies.

Standard Gaussian

We first test the Random Walk MH for the Gaussian with different tuning parameters σ in the proposal distribution.

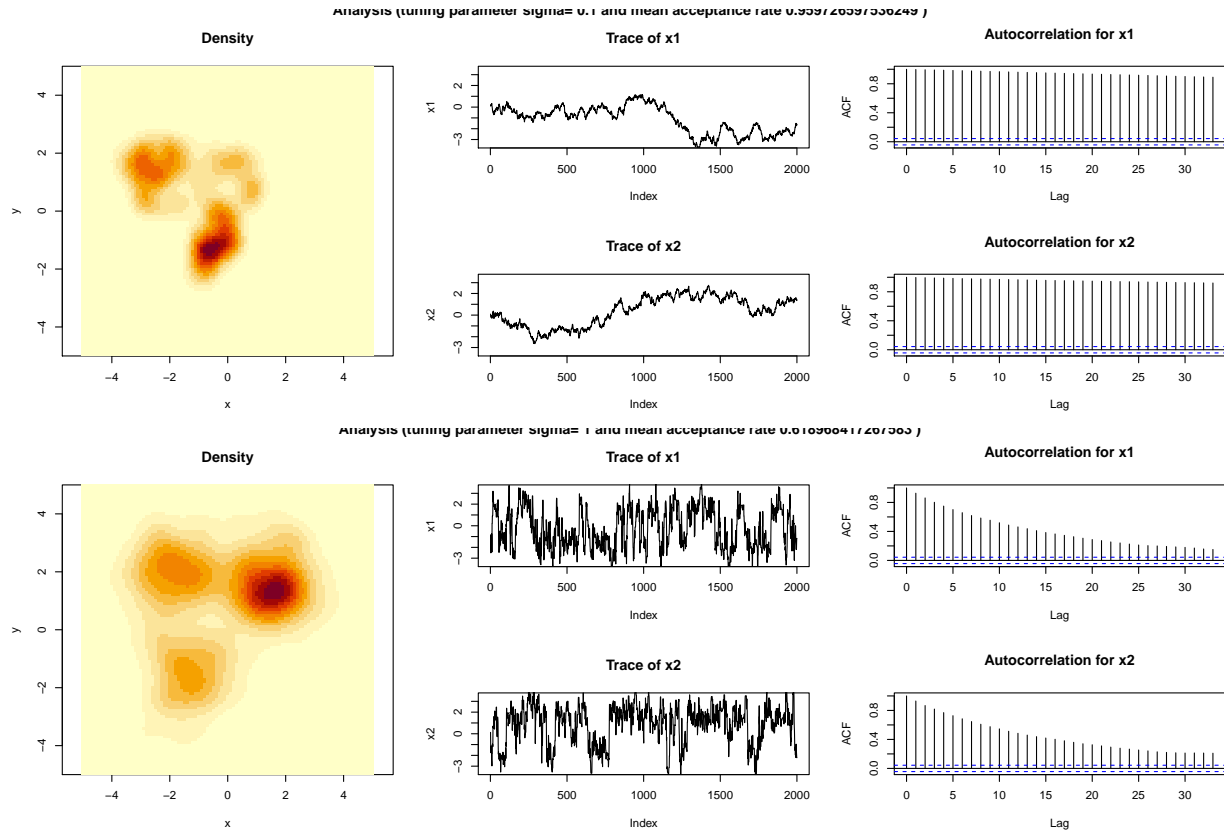


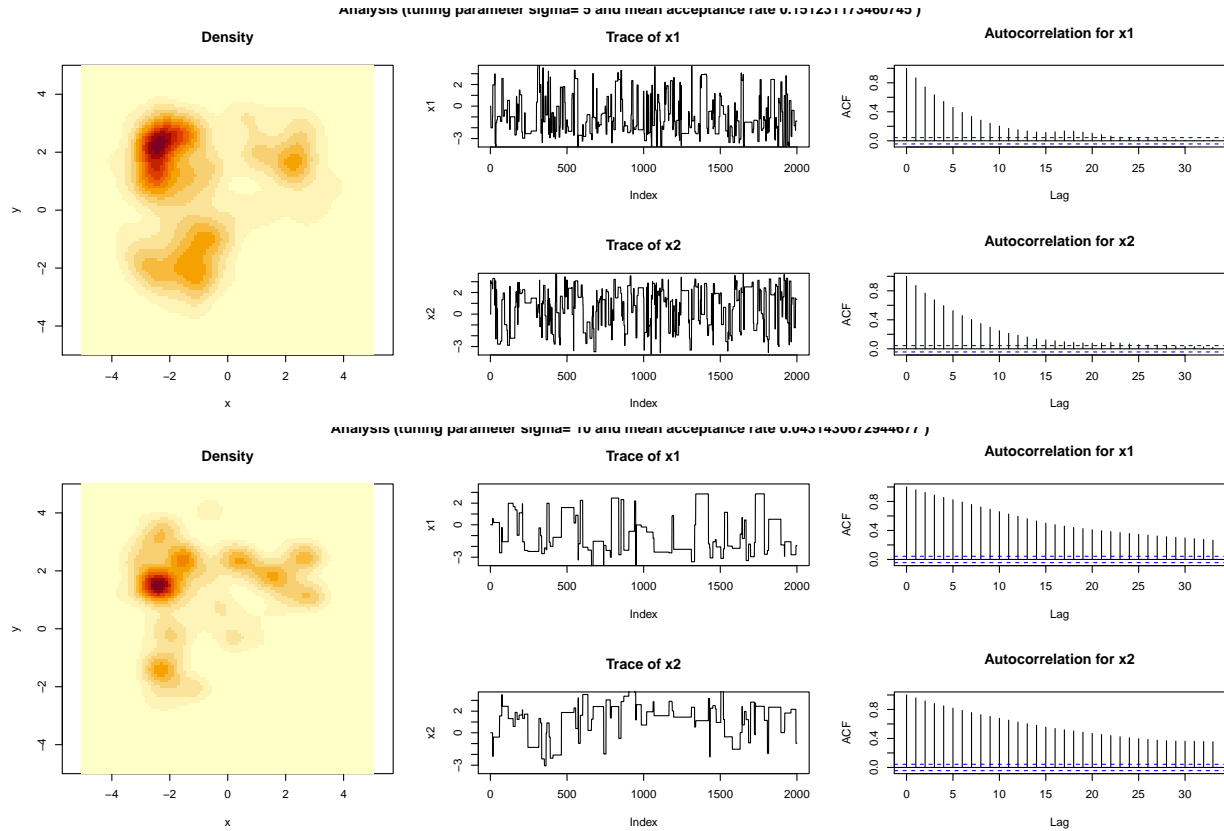


Depending on the choice of the tuning parameter the MCMC algorithms show different efficiency. Actually all tuning parameters explore the state-space rather sedately. However, for values $\sigma < 0.5$ the mean acceptance rate gets bigger than recommended and the traces change in too little step, for values $\sigma > 1.5$ the traces start to pause in a level due to a too small acceptance rate. We recommend values $\sigma \in (0.5, 1.5)$ for the Gaussian, since for those values the autocorrelation has the smallest lag and shrinks at least after 20 steps (what is still much).

Multimodal

We continue with application of the Random Walk MH to the Multimodal with different tuning parameters σ in the proposal distribution.

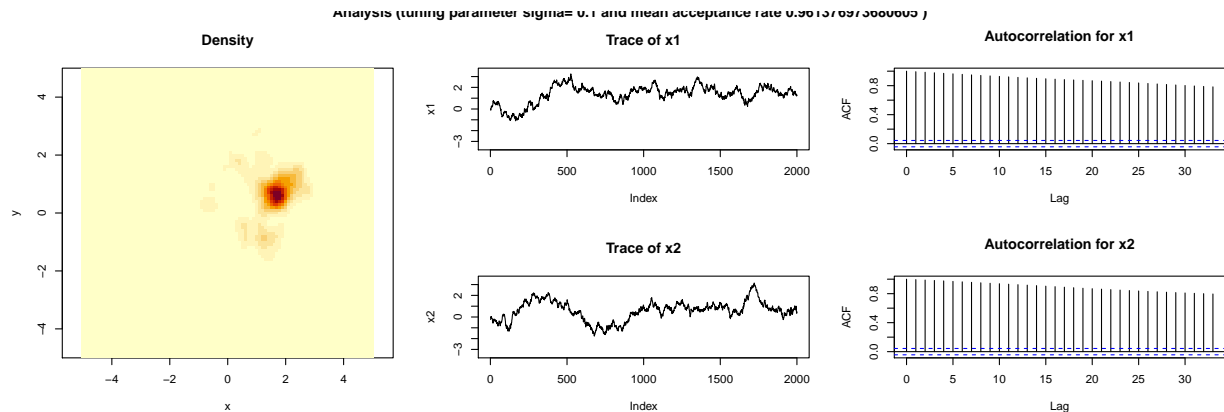


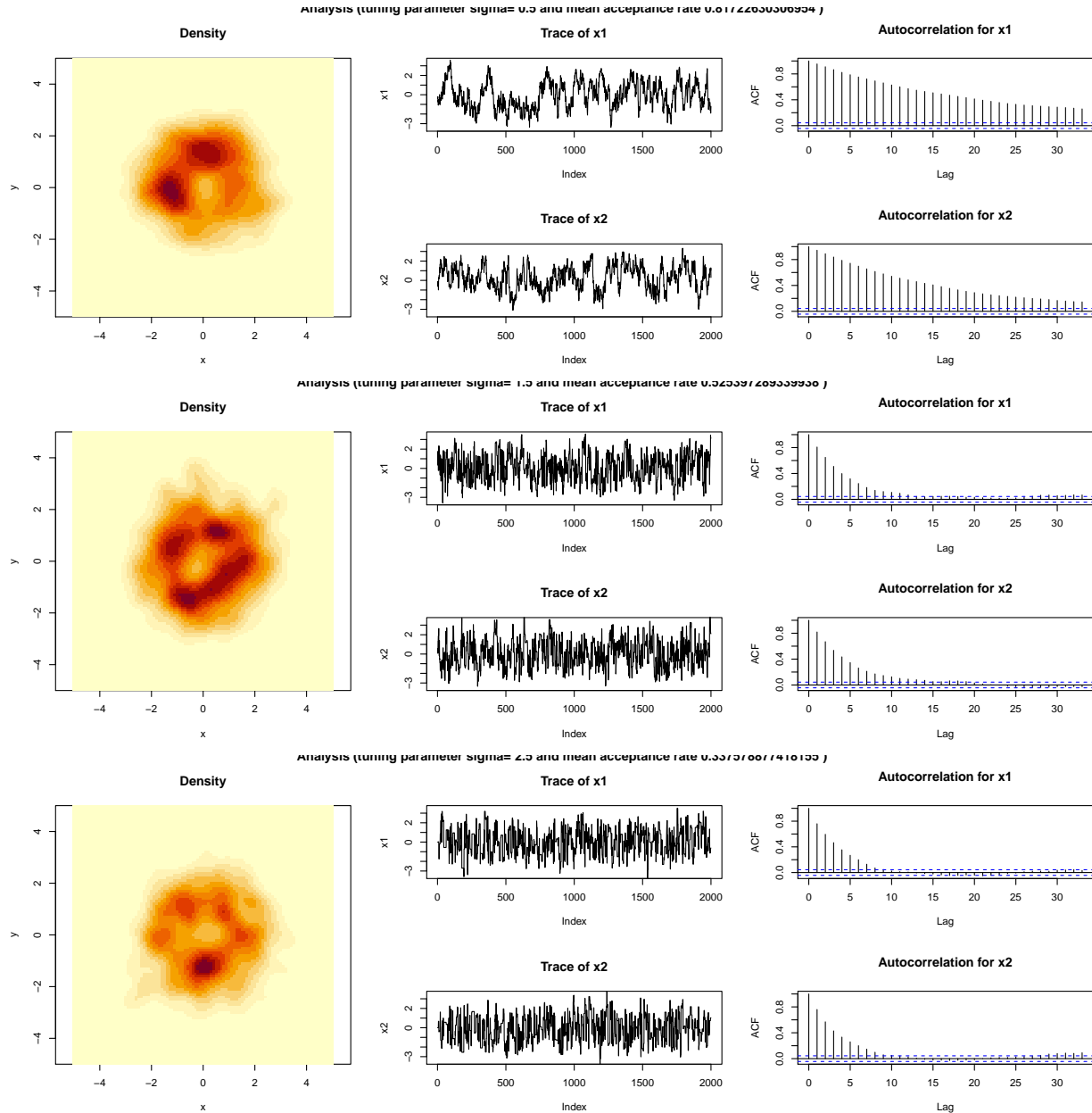


For the smallest choice of $\sigma = 0.1$ the chain does not really explore all modes and stays for quite long in one mode when it is there, therefore the autocorrelation is very high. When increasing up to $\sigma = 5.0$ we improve the exploration of the different modes and reduce the autocorrelation significantly. Moreover, the mean acceptance rate reaches the recommended range. For very high tuning parameter choices $\sigma = 10.0$, we start to wildly jump from one mode to the other, but the traces start to pause too long between the jumps.

Volcano

Finally, we try the random walk MH for the volcano shaped density.





If the tuning parameter is chosen too small $\sigma = 0.1$ the chain does not explore the entire ring, but get stuck. When increasing σ the chain starts to walk along the circle, but for higher parameters like $\sigma = 1.5$ the chain explores the ring with a short autocorrelation (figuratively speaking, it can also jump from one side toe the other and does not need to walk along the circle).

Conclusion

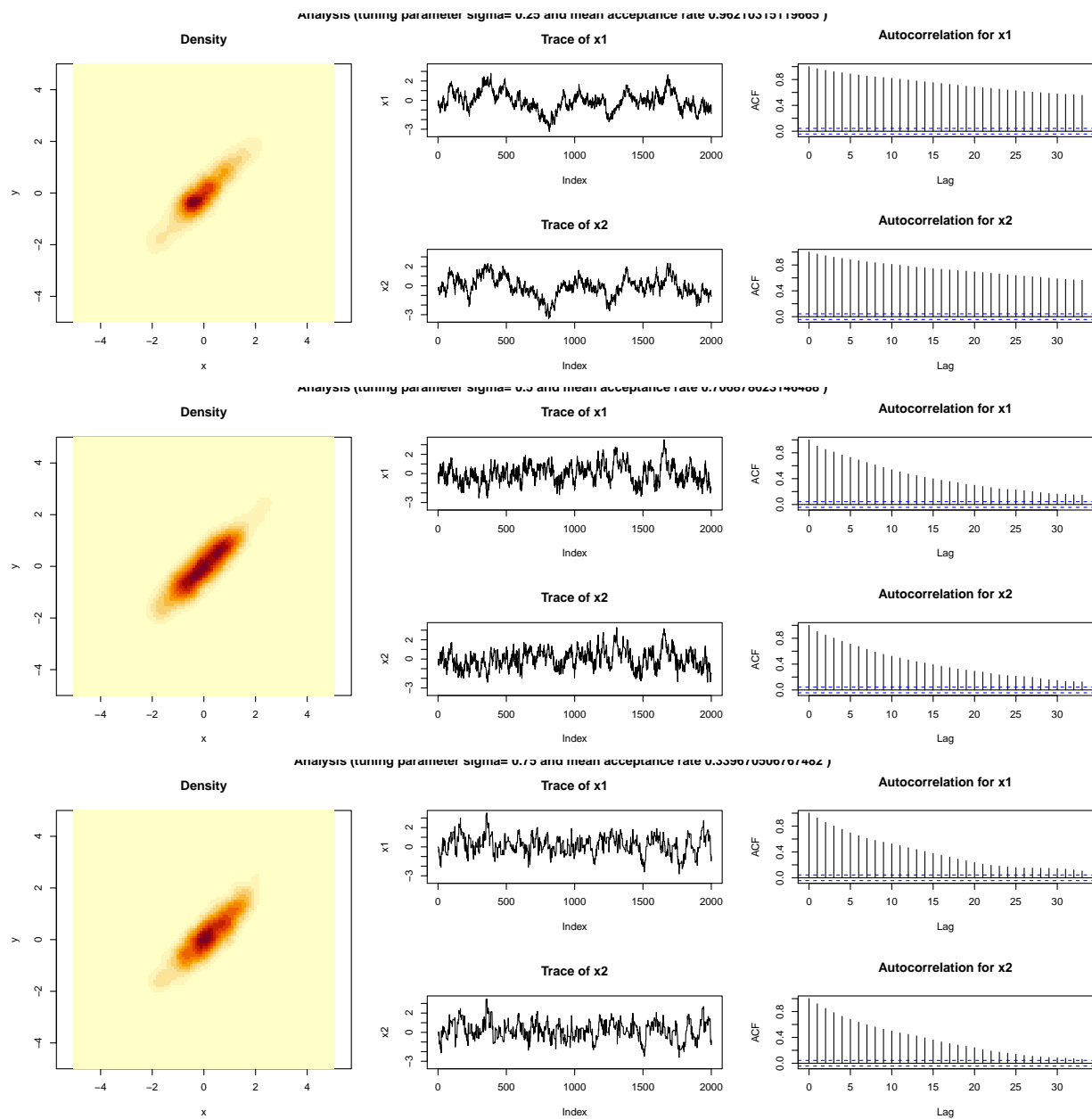
For all examples, there is a range of runing parameters which explore the state-space, but the autocorrelation is still pretty high.

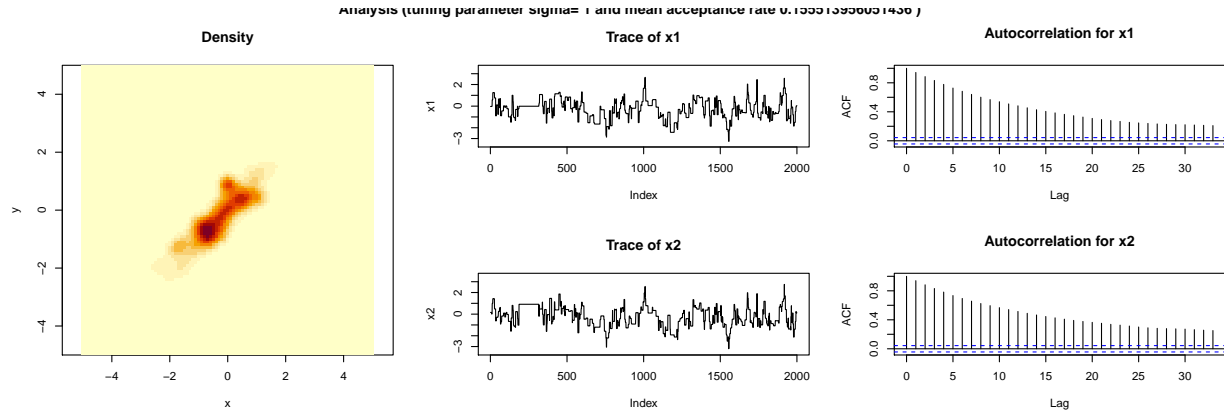
1.3 Langevin MH

Based on the Langevin dynamics and its Euler-Maruyama discretisation, the MALA algorithm uses gradient information to define the proposal density.

Standard Gaussian

We first test the Langevin MH for the Gaussian with different tuning parameters σ in the proposal distribution.

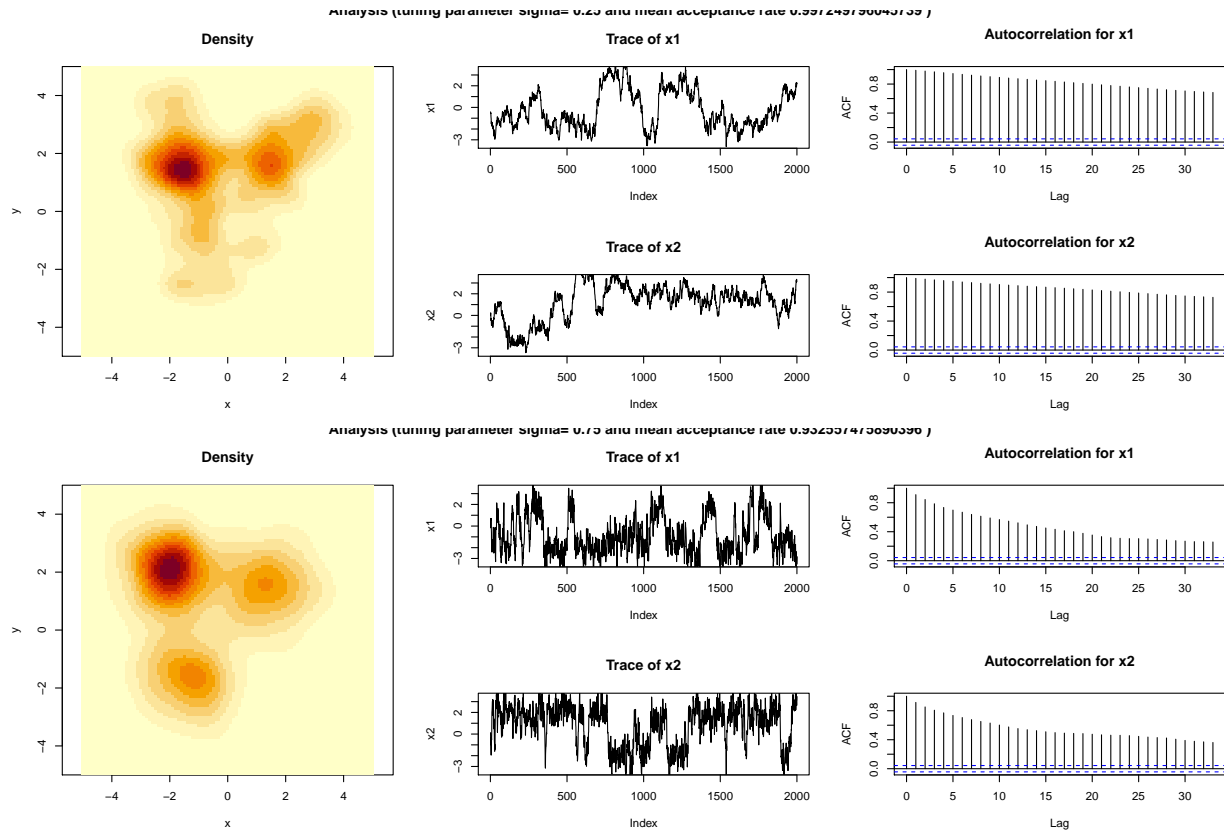


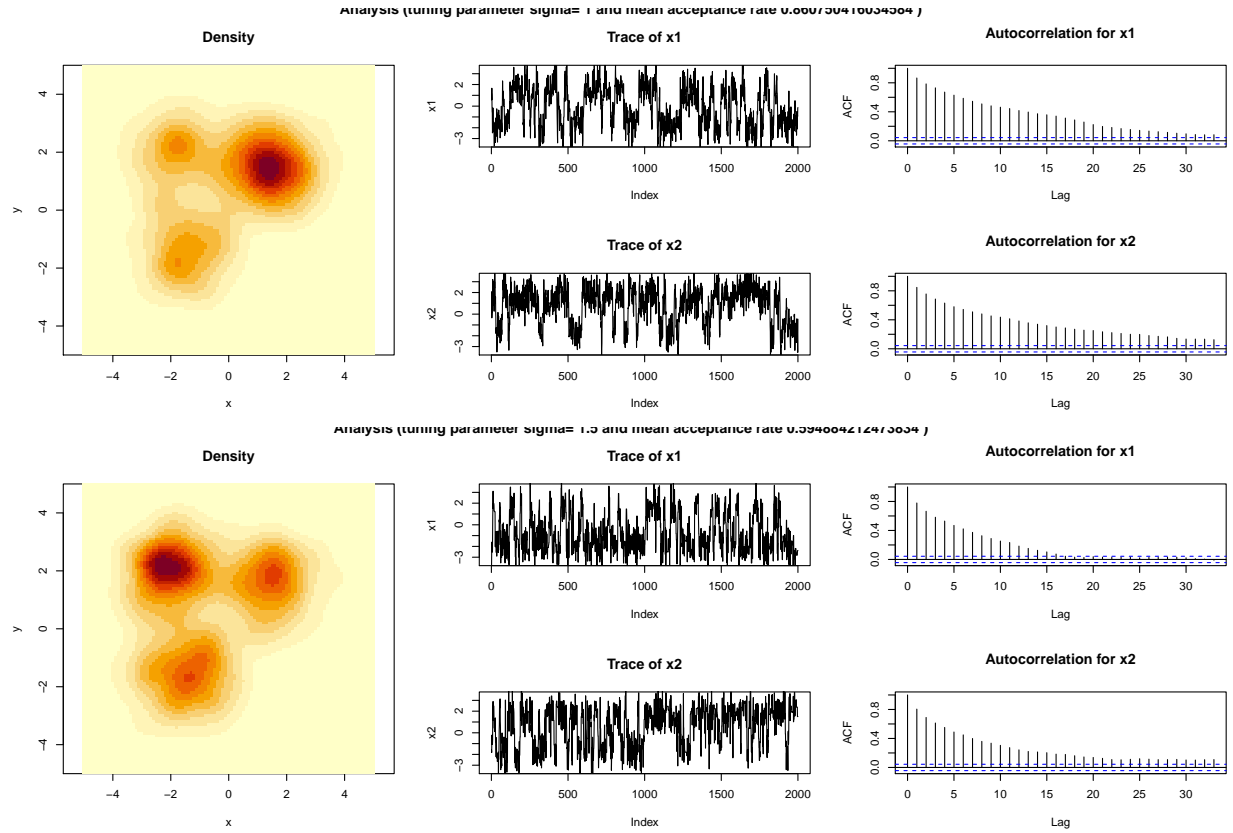


Even if the mean acceptance rate for small parameters $\sigma \leq 0.5$ is in the asymptotically optimal range, the trace evolves too slow and with too high correlation. For $\sigma = 0.75$ the target density is replicated - the autocorrelation decreases and the exploration is at least acceptable. Already for $\sigma = 1.0$ the acceptance rate is too small and the chain pauses too long to explore the space.

Multimodal

We continue with application of the Langevin MH to the Multimodal with different tuning parameters σ in the proposal distribution.

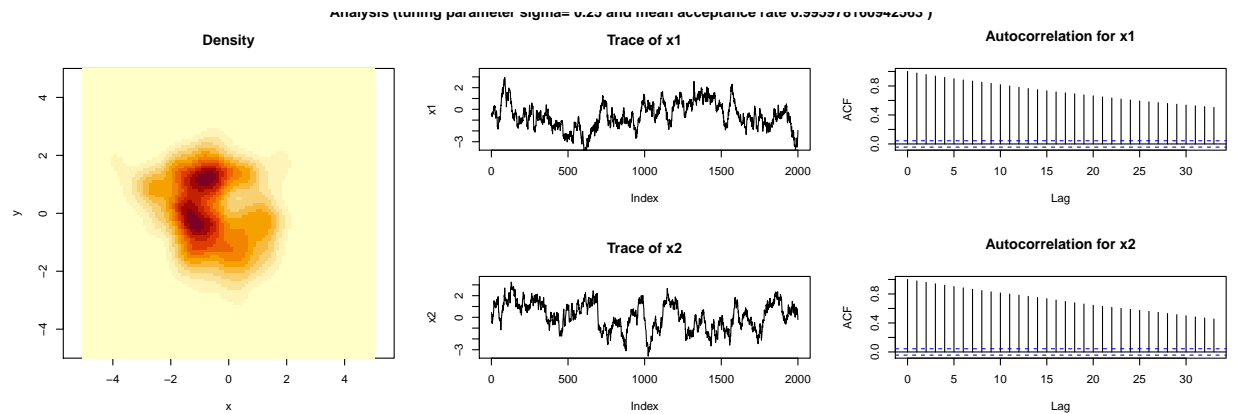


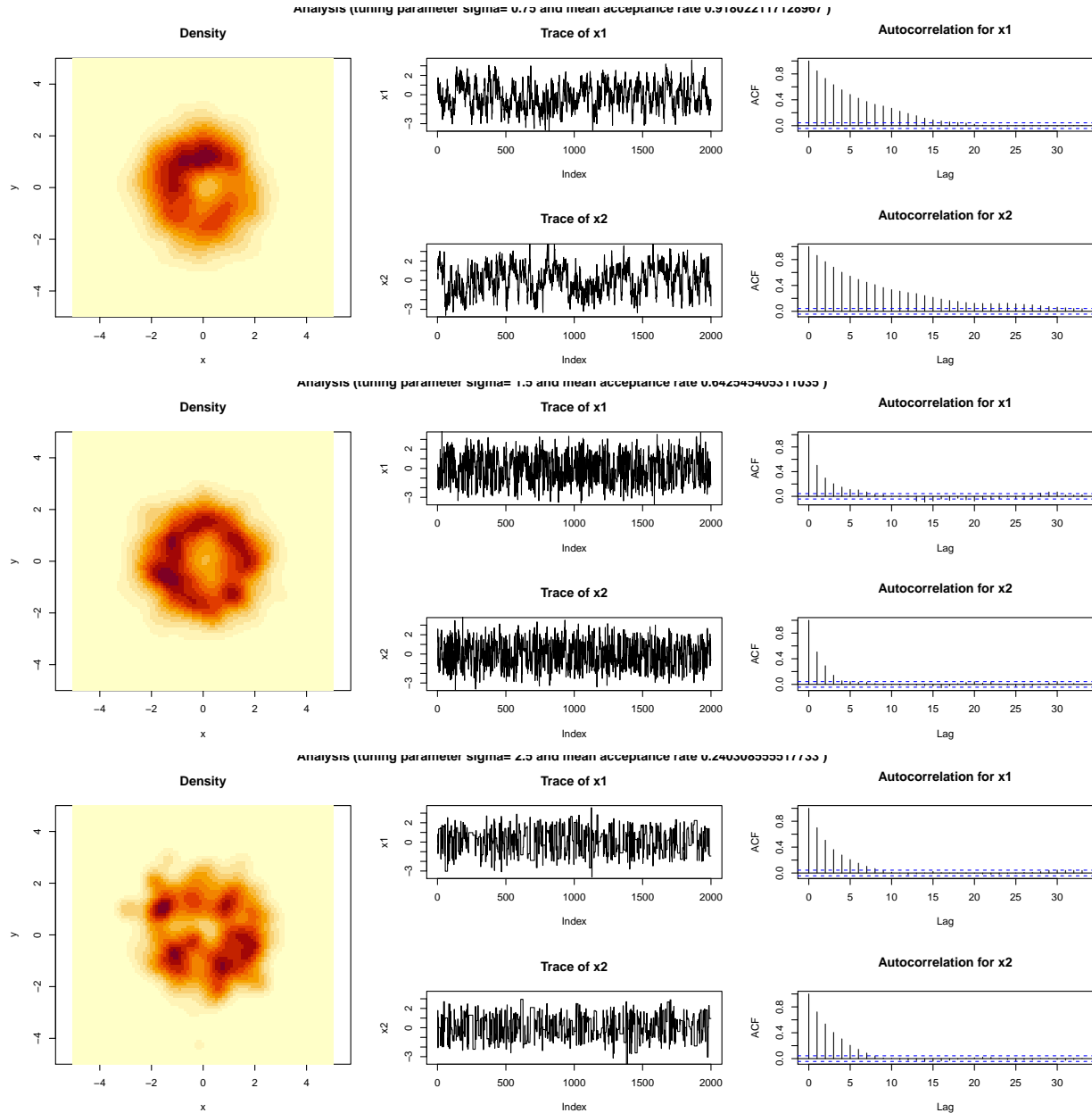


For small tuning parameter $\sigma \leq 0.75$ the Langevin chain get stuck in the local maxima of the multimodal dsitribution too long and hence does explore the other modes too badly. For higher parameters $\sigma \geq 1.0$ the chains reduce their autocorrelation and start to explore all modes.

Volcano

Finally, we try the Langevin MH for the volcano shaped density.





For reasonably big choices of $\sigma = 1.5$ a chain can be generated which does explore the entire ring with small autocorrelation and an acceptance rate close to the asymptotic optimum. For too small and too big choices the same as for the Random Walk holds: Only a fraction of the ring is explored for too small $\sigma = 0.25$, then the chains starts to walk along the circle for $\sigma = 0.5$. For too big proposal spread with $\sigma = 2.5$ too many large steps will be rejected.

Conclusion

We saw that the MALA algorithm is very sensitive to small changes in tuning parameter σ . For too small choices it tends to get stuck in local maxima, what is a known issue and we realized that behaviour here as well.

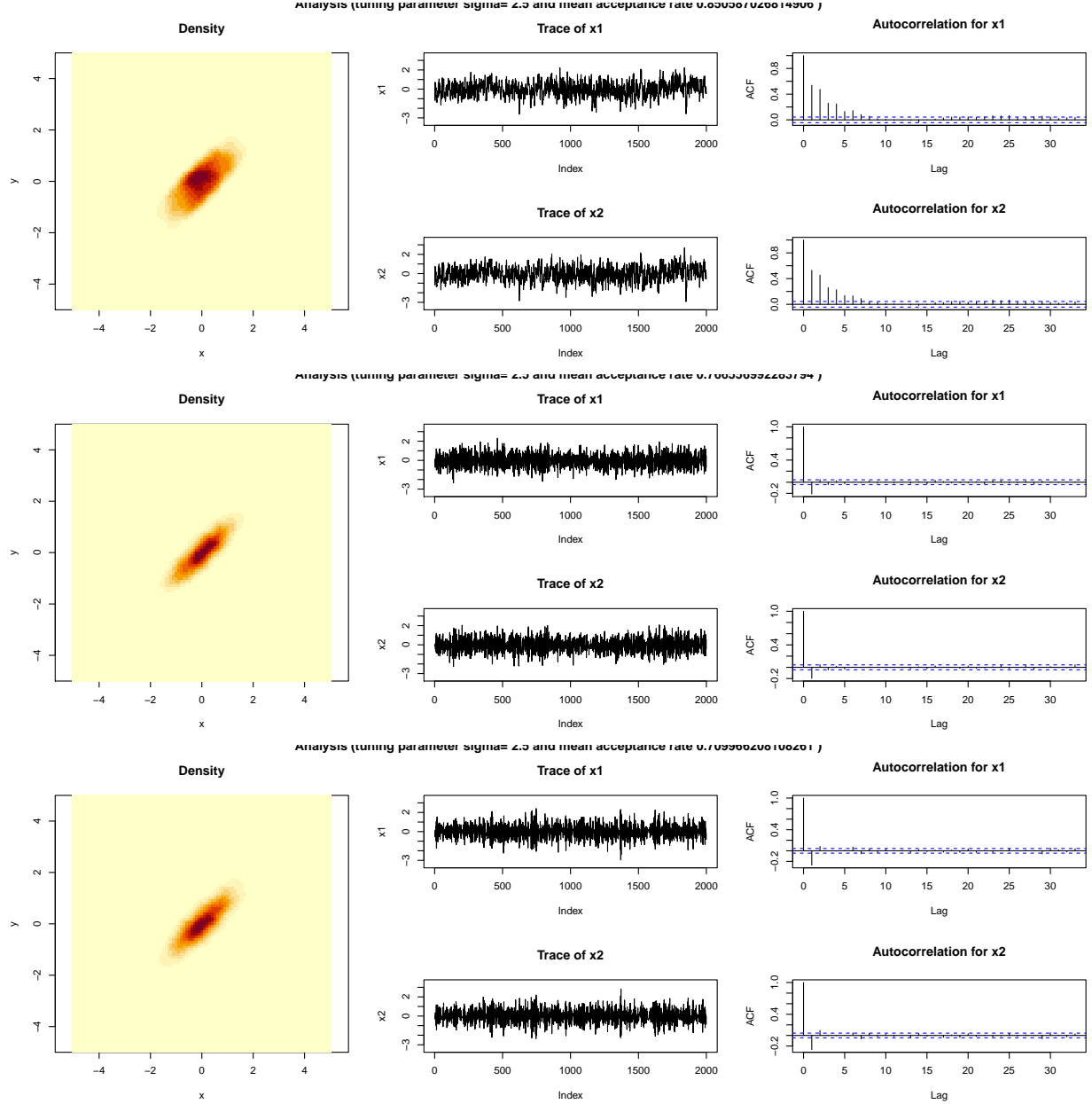
For the Gaussian distribution, the Langevin MCMC version works nicely, but the Random Walk does not perform bad here either. In the Multimodal case, the Random walk with suitable proposal spread does not get stuck in a maximum as the Langevin did. The circular density without clear mode is difficult for both

MCMC versions, but the Langevin can be parametrised to handle is better. However, in general it depends a lot on the proper tuning for both.

1.4 Hamiltonian MH

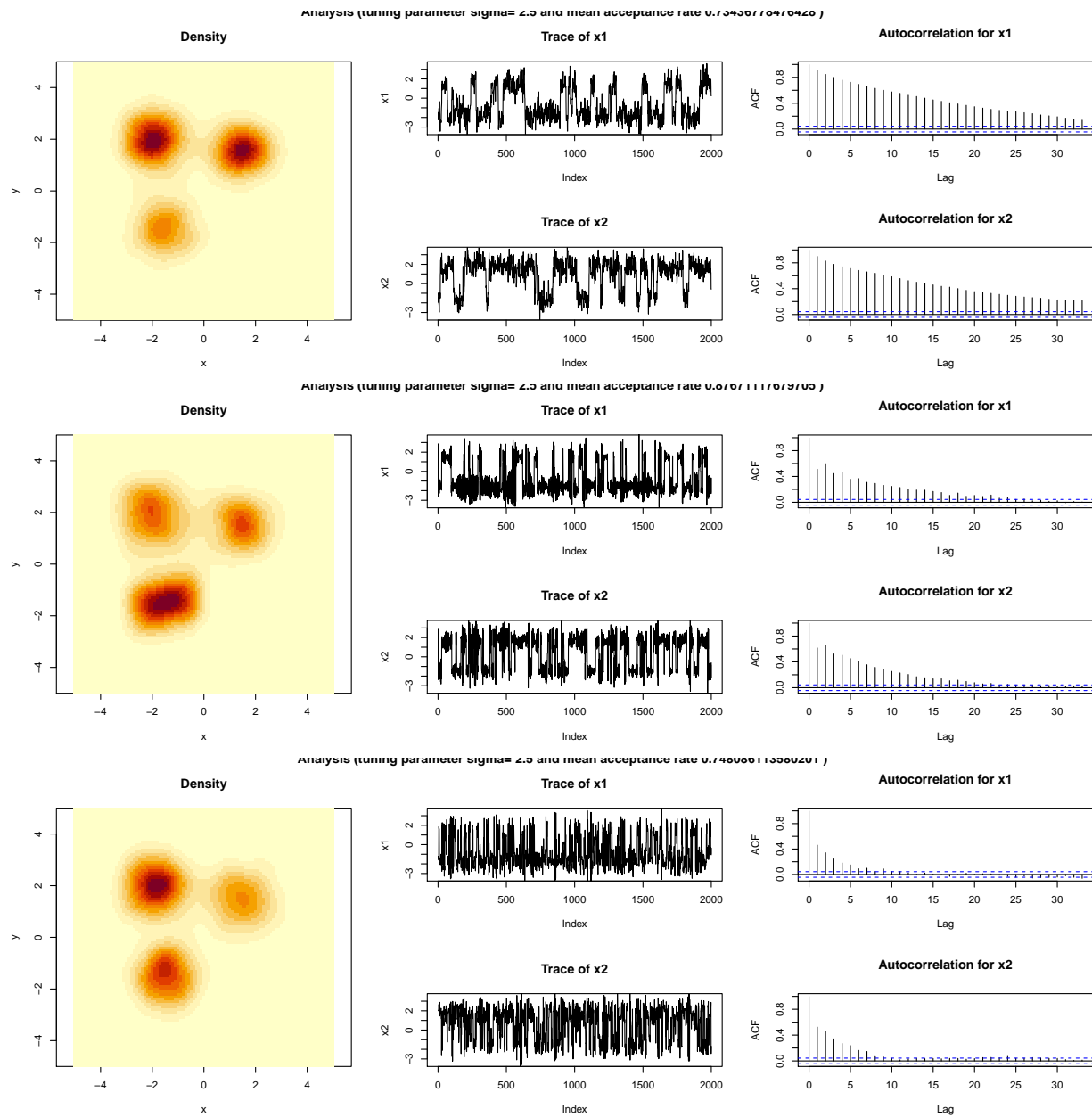
Standard Gaussian

We first test the Hamiltonian MH for the Gaussian with different tuning parameters σ in the proposal distribution.



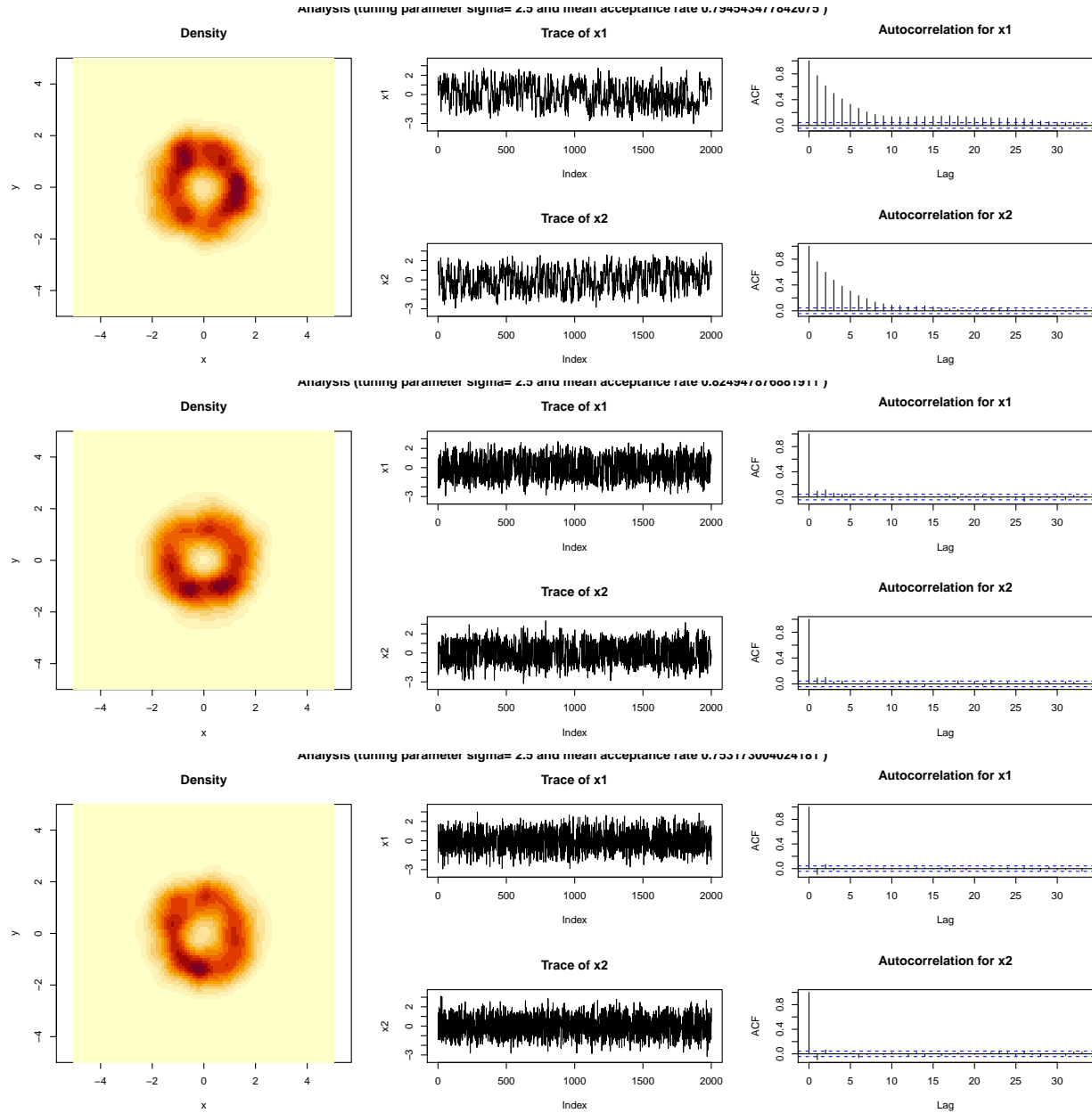
Multimodal

We continue with application of the Hamiltonian MH to the Multimodal with different tuning parameters σ in the proposal distribution.



Volcano

Finally, we try the Hamiltonian MH for the volcano shaped density.



Conclusion

- Burn-in
- All require tuning
- Compare performance

2 RStan

We consider the example of George et al. For a set of $N = 10$ pumps we observe the values - y_i : number of times that pump i failed - t_i : operation time of pump i whose data is used as input in the next chunk.

The numbers of failures per pump are modelled by a $Poisson(\lambda_i t_i)$ likelihood together with a $Gamma(\alpha, \beta)$ distributed prior for the λ_i with always $i = 1, \dots, N$. Additionally, the hyper-prior for α is $Exp(1.0)$ and for β it is $Gamma(0.1, 1.0)$ respectively. Therewith, we define a `stan` model.

NB! We do not use a separate `.stan`-file but generate the model named “pump” directly in the environment. The syntax for the model generation is exactly the same as in files, but later the function call for the model fit deviates! The `pump.stan` file is only included for the seek of accordance with the exercise sheet.

```
// generates a stan model named pump in the current environment

data{
  int<lower=0> N;      // number of pumps
  int<lower=0> y[N];   // number of failures
  real<lower=0> t[N];  // operation times of pumps
}

parameters{
  real<lower=0> lambda[N];
  real<lower=0> alpha;
  real<lower=0> beta;
}

transformed parameters{
  real<lower=0> eta[N];
  for (i in 1:N)
    eta[i] = lambda[i] * t[i];
}

model{
  target += exponential_lpdf( alpha | 1.0 );      // hyper-prior log-density
  target += gamma_lpdf( beta | 0.1, 1.0 );        // hyper-prior log-density
  target += gamma_lpdf( lambda | alpha, beta );   // prior log-density
  target += poisson_lpmf( y | eta );              // likelihood log-density
}
```

We sample from the posterior distribution using `stan`.

NB! The function call is different since we have already a stan model called “pump” in the cache and can sample from that directly.

```
## Loading required package: StanHeaders
## Loading required package: ggplot2
## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
## Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file
##
## SAMPLING FOR MODEL '2834f5355f713513a0207a56538df8ff' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
```

```

## Chain 1: Iteration: 200 / 2000 [ 10%] (Warmup)
## Chain 1: Iteration: 400 / 2000 [ 20%] (Warmup)
## Chain 1: Iteration: 600 / 2000 [ 30%] (Warmup)
## Chain 1: Iteration: 800 / 2000 [ 40%] (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%] (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 0.07 seconds (Warm-up)
## Chain 1: 0.064 seconds (Sampling)
## Chain 1: 0.134 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL '2834f5355f713513a0207a56538df8ff' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration: 1 / 2000 [ 0%] (Warmup)
## Chain 2: Iteration: 200 / 2000 [ 10%] (Warmup)
## Chain 2: Iteration: 400 / 2000 [ 20%] (Warmup)
## Chain 2: Iteration: 600 / 2000 [ 30%] (Warmup)
## Chain 2: Iteration: 800 / 2000 [ 40%] (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%] (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 2:
## Chain 2: Elapsed Time: 0.063 seconds (Warm-up)
## Chain 2: 0.082 seconds (Sampling)
## Chain 2: 0.145 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL '2834f5355f713513a0207a56538df8ff' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 0 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration: 1 / 2000 [ 0%] (Warmup)
## Chain 3: Iteration: 200 / 2000 [ 10%] (Warmup)
## Chain 3: Iteration: 400 / 2000 [ 20%] (Warmup)
## Chain 3: Iteration: 600 / 2000 [ 30%] (Warmup)
## Chain 3: Iteration: 800 / 2000 [ 40%] (Warmup)

```

```

## Chain 3: Iteration: 1000 / 2000 [ 50%] (Warmup)
## Chain 3: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 3: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 3: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 3: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 3: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 3: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 3:
## Chain 3: Elapsed Time: 0.081 seconds (Warm-up)
## Chain 3: 0.079 seconds (Sampling)
## Chain 3: 0.16 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL '2834f5355f713513a0207a56538df8ff' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 0 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration: 1 / 2000 [ 0%] (Warmup)
## Chain 4: Iteration: 200 / 2000 [ 10%] (Warmup)
## Chain 4: Iteration: 400 / 2000 [ 20%] (Warmup)
## Chain 4: Iteration: 600 / 2000 [ 30%] (Warmup)
## Chain 4: Iteration: 800 / 2000 [ 40%] (Warmup)
## Chain 4: Iteration: 1000 / 2000 [ 50%] (Warmup)
## Chain 4: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 4: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 4: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 4: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 4: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 4: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 4:
## Chain 4: Elapsed Time: 0.072 seconds (Warm-up)
## Chain 4: 0.06 seconds (Sampling)
## Chain 4: 0.132 seconds (Total)
## Chain 4:

```

Finally, we investigate the results of the stan fit.

```

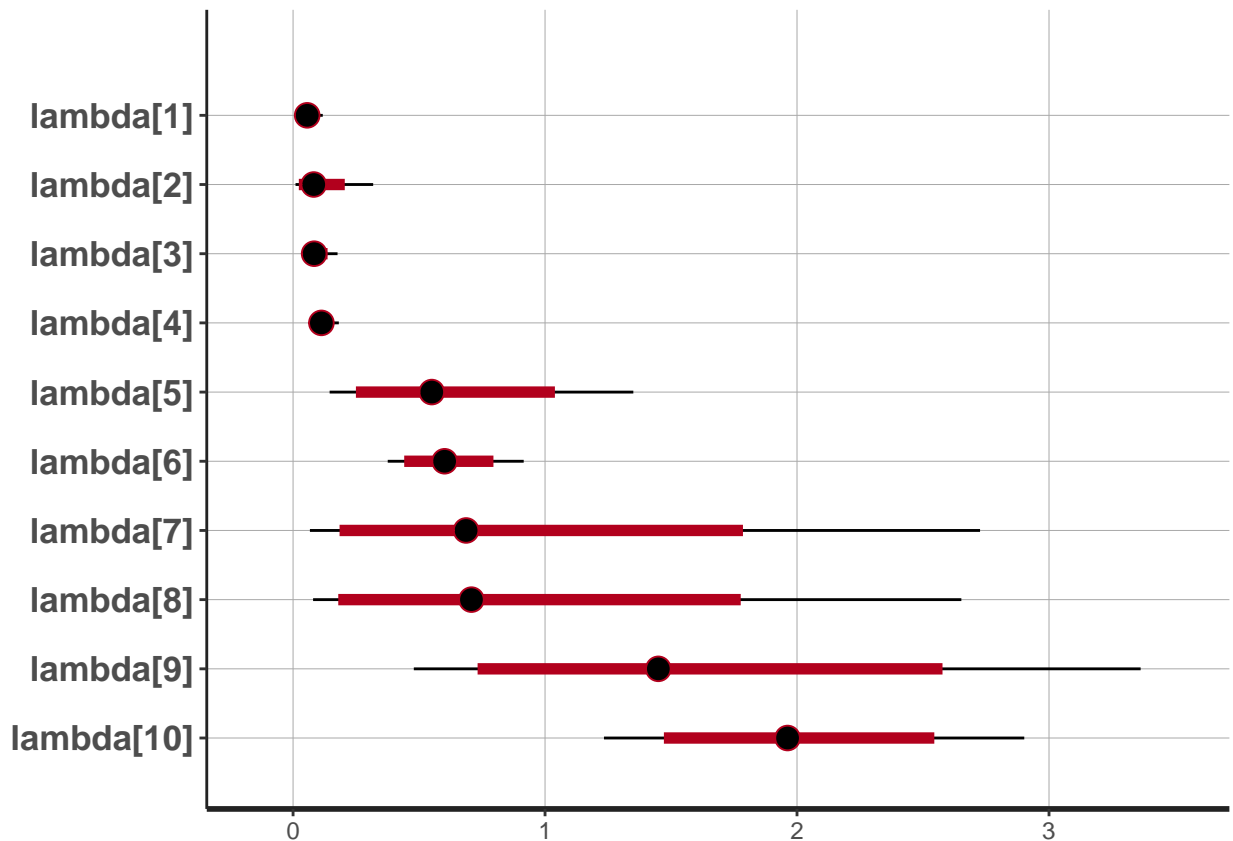
## Inference for Stan model: 2834f5355f713513a0207a56538df8ff.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
## lambda[1]	0.06	0.00	0.03	0.02	0.04	0.06	0.07	0.12	5258	1
## lambda[2]	0.10	0.00	0.08	0.01	0.04	0.08	0.14	0.32	5481	1
## lambda[3]	0.09	0.00	0.04	0.03	0.06	0.08	0.11	0.18	5365	1
## lambda[4]	0.12	0.00	0.03	0.06	0.09	0.11	0.14	0.18	5887	1
## lambda[5]	0.60	0.00	0.32	0.15	0.37	0.55	0.78	1.35	4665	1
## lambda[6]	0.61	0.00	0.14	0.38	0.51	0.60	0.69	0.92	5004	1
## lambda[7]	0.88	0.01	0.72	0.07	0.37	0.69	1.20	2.73	4797	1
## lambda[8]	0.87	0.01	0.70	0.08	0.38	0.71	1.17	2.65	5328	1
## lambda[9]	1.58	0.01	0.75	0.48	1.04	1.45	2.00	3.36	4642	1
## lambda[10]	1.99	0.01	0.42	1.23	1.69	1.96	2.26	2.90	5228	1

```

## alpha      0.70    0.01 0.28    0.28    0.49    0.65    0.85    1.35 2693    1
## beta       0.93    0.01 0.57    0.18    0.52    0.83    1.23    2.33 3100    1
## eta[1]     5.61    0.03 2.36    1.97    3.92    5.29    6.98   11.12 5258    1
## eta[2]     1.60    0.02 1.23    0.15    0.70    1.29    2.17    4.99 5481    1
## eta[3]     5.56    0.03 2.33    2.02    3.85    5.27    6.87   11.10 5365    1
## eta[4]    14.58    0.05 3.88    8.03   11.79   14.18   17.07   22.89 5887    1
## eta[5]     3.16    0.02 1.67    0.76    1.93    2.88    4.10    7.08 4665    1
## eta[6]    19.21    0.06 4.36   11.80   16.16   18.89   21.82   28.74 5004    1
## eta[7]     0.92    0.01 0.75    0.07    0.39    0.72    1.26    2.86 4797    1
## eta[8]     0.92    0.01 0.73    0.08    0.40    0.74    1.23    2.78 5328    1
## eta[9]     3.32    0.02 1.57    1.01    2.19    3.04    4.19    7.06 4642    1
## eta[10]    20.88    0.06 4.44   12.96   17.70   20.61   23.69   30.47 5228    1
## lp__      -43.98    0.07 2.59  -49.62  -45.53  -43.60  -42.13  -39.99 1467    1
##
## Samples were drawn using NUTS(diag_e) at Wed Feb 17 15:37:10 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
## ci_level: 0.8 (80% intervals)
## outer_level: 0.95 (95% intervals)

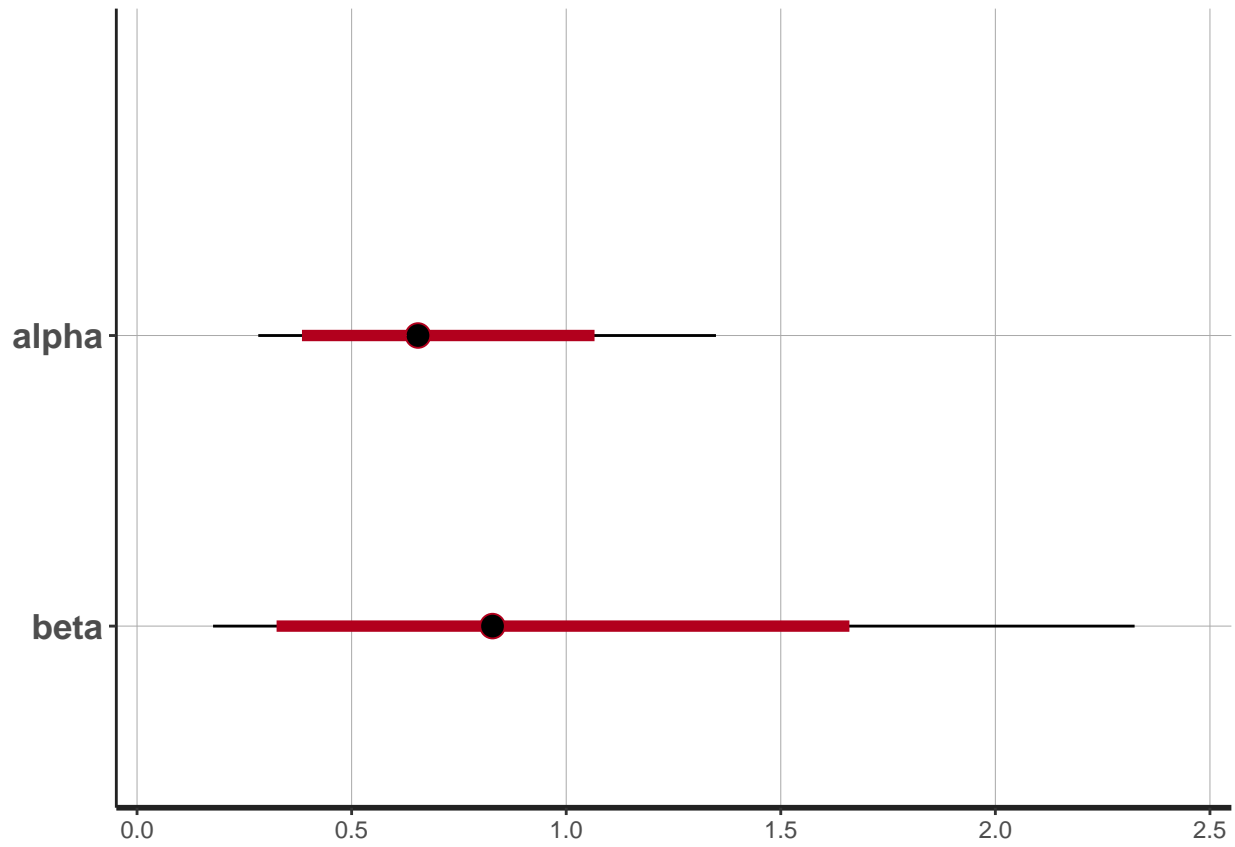
```



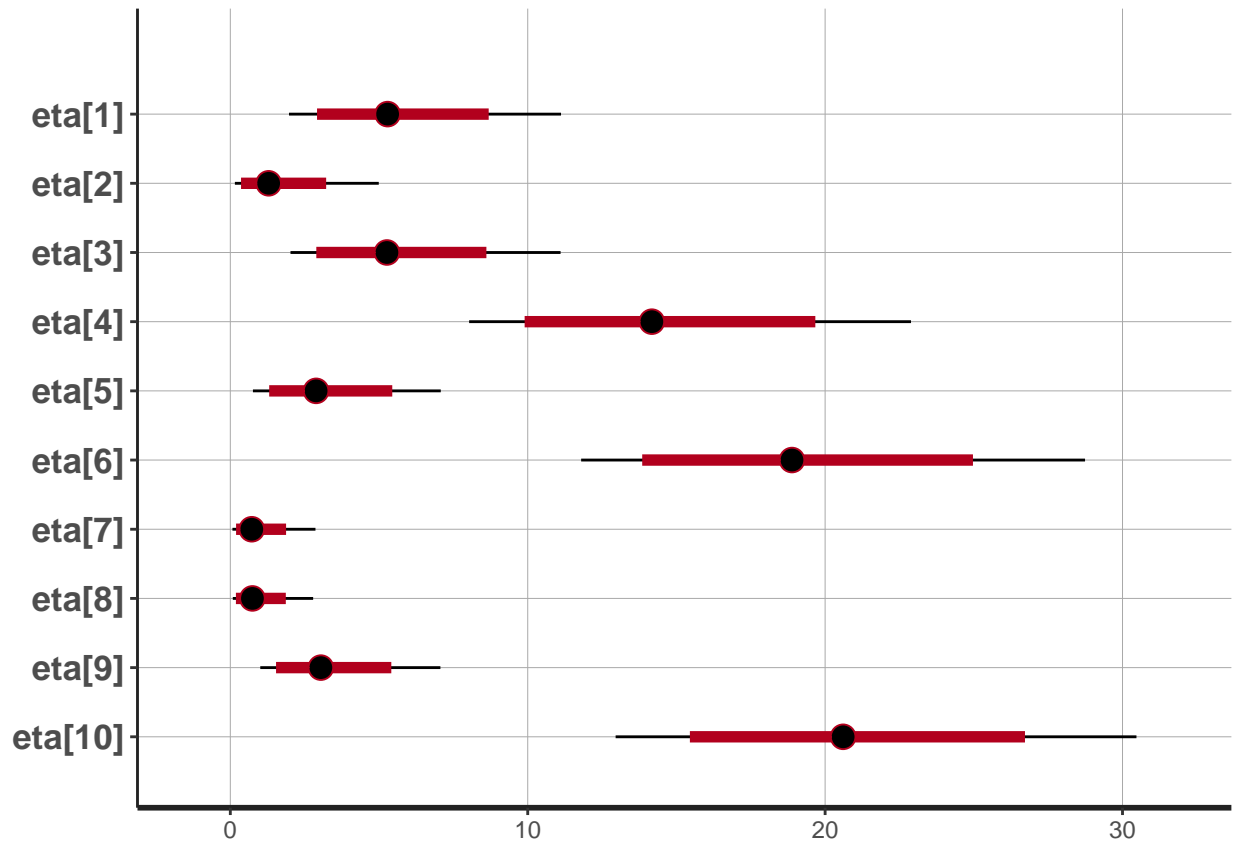
```

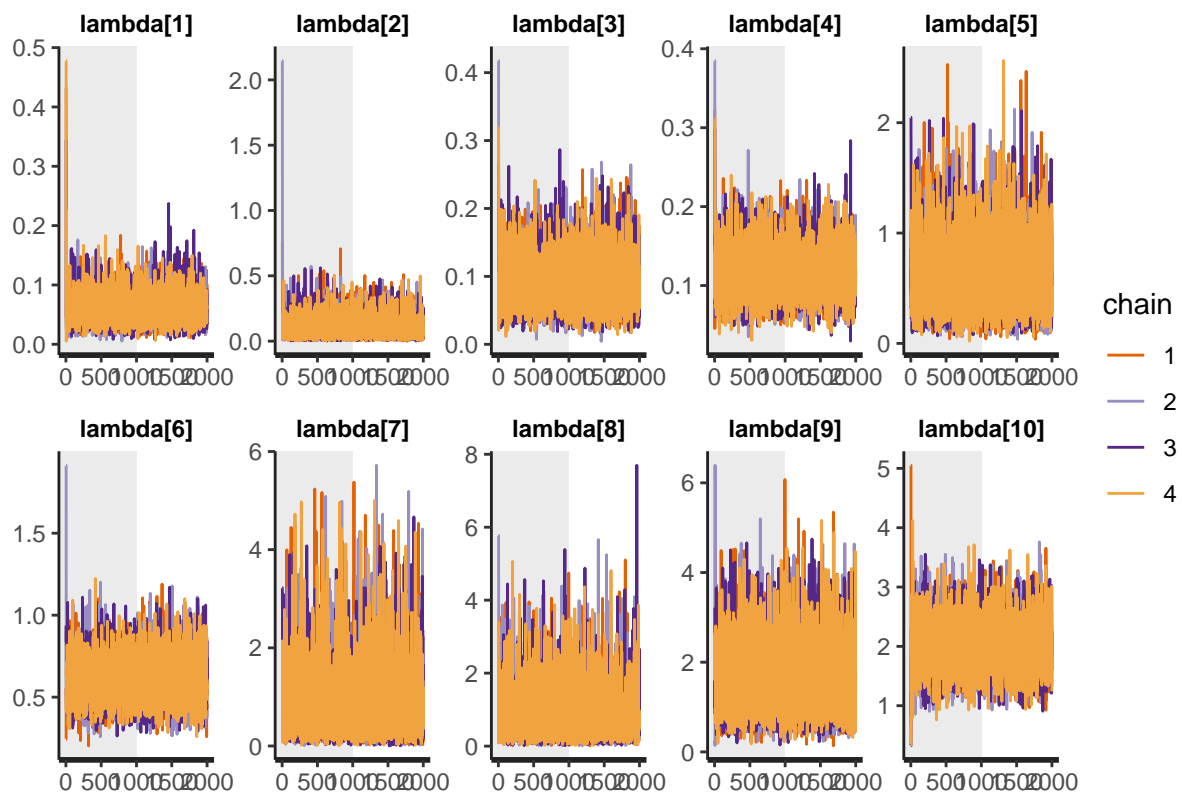
## ci_level: 0.8 (80% intervals)
## outer_level: 0.95 (95% intervals)

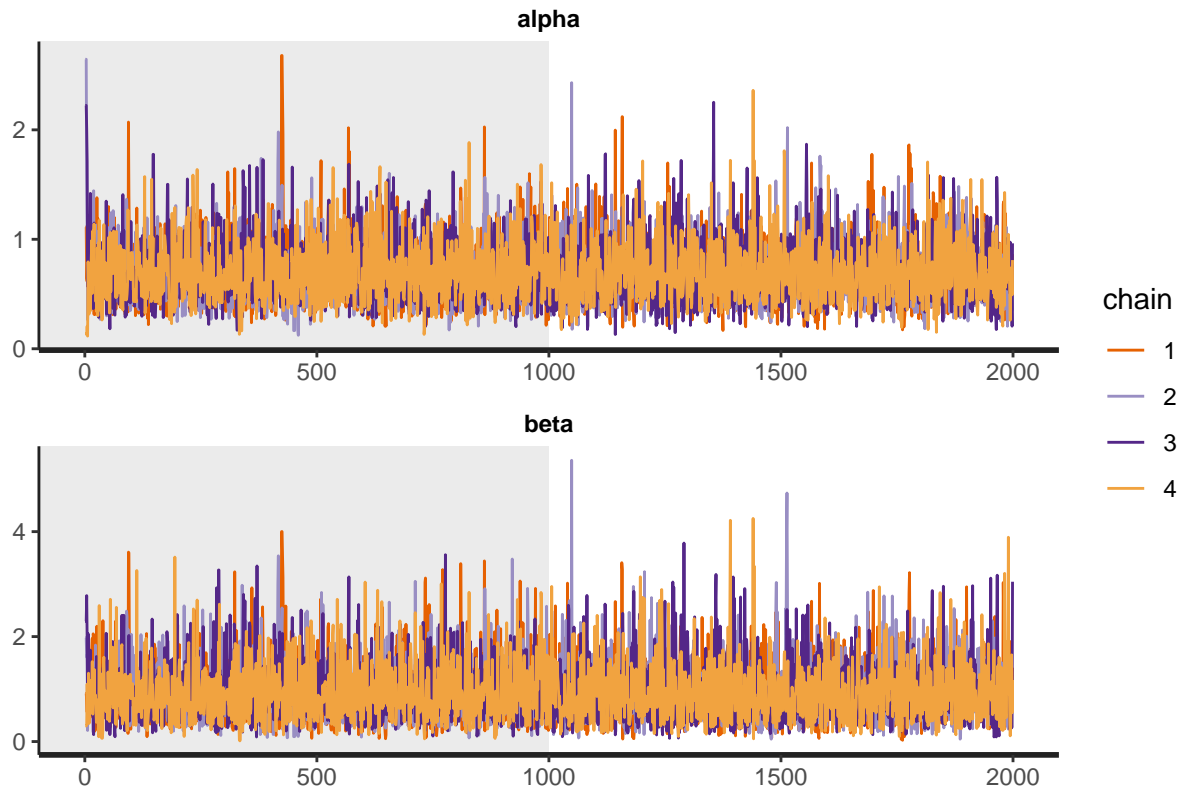
```

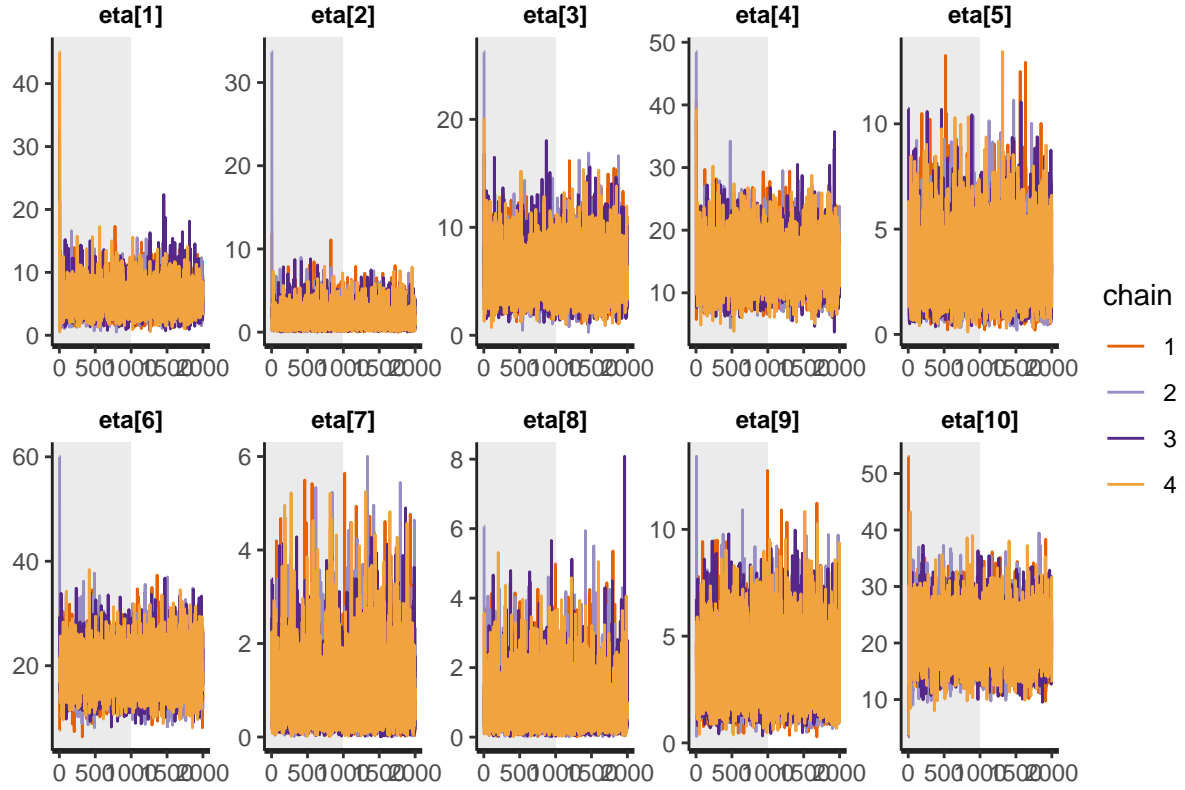



```
## ci_level: 0.8 (80% intervals)
## outer_level: 0.95 (95% intervals)
```









From the print output of the fitted model, we analyse the effective sample size n_{eff} . With only 4.000 post-warmup draws in total, for the λ (and η) the effective samples size is around 5.000 and thereby surprisingly higher as the actual number of samples! In contrast for α and β , the effective sample size is only around 2500 and 3000, respectively, such that the quality of those samples is not as good as for the oother quantities.

In the trace plots for λ (and η) we see that the state space is explored very effectively throughout the entire chains. Maybe for $i \in \{5, 7, 9\}$ the range of lower probability is not perfectly covered, but that is hard/impossible to say without further histograms. In general, we can assume that the chains reached a behaviour of the limit distribution already after a few dozen steps. For α and β we observe some higher autocorrelation and a less effective exploration of the target space. For a few dozen steps the sampels appear to remain close before the next jump comes, what is an undesired behavior.