Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci

Susmita Sridar [1 †] ✉, **Mathias Spliid Heltberg** [2 †] ✉, **Christian Stentoft Michelsen** [2 †], **Judith Mine Hattab** [1], **Angela Taddei** [1]

[1]Institut Curie, PSL University, Sorbonne Universite, CNRS, Nuclear Dynamics, Paris, France; [2]Niels Bohr Institute, University of Copenhagen

✉ **For correspondence:** mathias.heltberg@nbi.ku.dk (MH)

[†]Authors contributed equally.

Present address: Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

Data availability: Data availability is available on Zenodo or the Github repository.

Funding: This work was supported by XXX Foundation. The funders had no role in the decision to publish.

Competing interests: The author declare no competing interests.

Abstract

In order to obtain fine-tuned regulation of protein production while maintaining cell integrity, it is of fundamental importance to living organisms to express a specific subset of the genes abailable in the genome. One way to achieve this is through the formation of subcompartments in the nucleus, known as foci, that can form at various locations on the DNA fibers and repress the transcriptional activity of all genes covered. In this work we investigate the physical nature of such foci, by applying single molecule microscopy in living cells. Here we study the motion of the protein SIR3. By combining various statistical methods, and combining a frequentist with a bayesian approach, we extract the diffusion properties for motion in a repair foci. In order to obtain useful information based on this, we derive similar measures for the foci itself, the motion of SIR3 outside the foci and other mutants of the cell. We reveal that the behaviour inside a repair foci is highly immobile and we compare this to theoretical expressions. Based on this we hypothesize that the repair foci is probably not a result of a second order liquid-liquid phase separation but rather a so-called Polymer Bridgng Model with numerous binding sites.

1 | INTRODUCTION

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris. Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

2 | METHODS & MATERIALS

2.1 | Diffusion model

For each of the different types of data (XXX), we load in the cells and group them by cell number and ID. For each group we compute the distance Δr between the subsequent observations $\vec{x}_i$:

$$\Delta r_i = \|\vec{x}_{i+1} - \vec{x}_i\|. \tag{1}$$

E.g., for Wild Type 1, we find 914 groups across 43 different cells, leading to a total of $N = 10.025$ distances. We model the diffusion distances with a Rayleigh likelihood, where the Rayleigh distribution is given by:

$$\text{Rayleigh}(x;\sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x > 0. \tag{2}$$

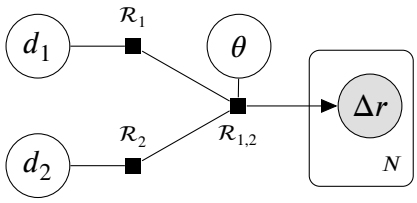


Figure 1. A graphical representation of the Bayesian model case of two diffusion components using the directed factor graph notation (Dietz, 2022). Here d_1 is the diffusion coefficient, $\mathcal{R}_1$ is the d-parameterized Rayleigh distribution and $\mathcal{R}_{1,2}$ is the mixture model of the Rayleigh distributions with a θ prior.

In this study, we parameterize the Rayleigh distribution in terms of the diffusion coefficient d, which is related to the scale parameter σ in eq. (2), through the XXX parameter, τ:

$$\sigma = \sqrt{2d\tau}, \tag{3}$$

with $\tau = 0.02$ in the current study. In the simplest form, where we assume only a single diffusion coefficient, d, the Bayesian model for this process is:

$$\begin{aligned}
&[d \text{ prior}] && d \sim \text{Exponential}(0.1) \\
&[\text{transformation}] && \sigma = \sqrt{2d\tau} \\
&[\text{likelihood}] && \Delta r_i \sim \text{Rayleigh}(\sigma).
\end{aligned} \tag{4}$$

A more realistic diffusion model include more than a single diffusion coefficient. Figure 1 shows this for the two-component case in directed factor graph notation (Dietz, 2022). In particular, the figure shows the combination of the $K = 2$ diffusion coefficients d_k through a mixture model $\mathcal{R}_{1,2}$ of the two d-parameterized Rayleigh distributions $\mathcal{R}_k$ with a θ-prior. We model each of the distances as independent, indicated by the N-replications plate. In equations, the figure is similar to:

$$\begin{aligned}
&[d_1 \text{ prior}] && d_1 \sim \text{Exponential}(0.1) \\
&[d_2 \text{ prior (ordered)}] && d_2 \sim \text{Exponential}(0.1), \quad d_1 < d_2 \\
&[\bar{\theta} \text{ prior}] && \theta_1 \sim \text{Uniform}(0,1), \quad \bar{\theta} = \left[\theta_1, 1-\theta_1\right] \\
&[\text{mixture model}] && \mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}) = \text{MixtureModel}\left(\left[\mathcal{R}(d_1), \mathcal{R}(d_2)\right], \bar{\theta}\right) \\
&[\text{likelihood}] && \Delta r_i \sim \mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}).
\end{aligned} \tag{5}$$

2.2 | Model comparison

We can generalize the $K = 2$ diffusion model to higher values of K by having $d_1, \ldots, d_K$ (ordered such that $d_1 < d_k < d_K$ to prevent the classical label-switching problem in the case of mixture mod-

els (McLachlan and Peel, 2004)) diffusion coefficients and letting the mixture model's $\bar{\theta}$-prior be a random variable from a flat Dirichlet distribution (such that $\sum_k \theta_k = 1$). We find that including up to three diffusion coefficients yields appropriate results. To compare the three models of different complexity, we compute the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) which is a generalized version of the Akaike information criterion (AIC) useful for Bayesian model comparison (Gelman, Hwang, and Vehtari, 2014). In short, the WAIC is an approximation of the out-of-sample performance of the model and consists of two terms, the log-pointwise-predictive-density, lppd, and the effective number of parameters p_{WAIC}:

$$\mathrm{WAIC} = -2\left(\mathrm{lppd} - p_{\mathrm{WAIC}}\right). \tag{6}$$

The lppd is the Bayesian version of the accuracy of the model and p_{WAIC} is a penalty term related to the risk of over-fitting; complex models (usually) have higher values of p_{WAIC} than simple models, (McElreath, 2020). The minus 2 factor is just a scaling included for historical reasons leading to low WAICs being better. Given two models, A and B, we compute both the individual WAIC values, W_A and W_B, their standard deviations, σ_{W_A} and σ_{W_B}, their difference, $\Delta_{A,B}$, and the standard error of their difference, $\sigma_{\Delta_{A,B}}$.

2.3 | MSD and energy

After choosing the optimal model, we extract the slow diffusion coefficient from the model, d_{slow}, and use this to compute the mean squared displacement (MSD) for the groups with a mean diffusion $D = \langle \frac{\Delta r^2}{4\tau} \rangle$ being slow, where slow is defined as $D < d_{\mathrm{slow}} + 3\sigma_{\mathrm{slow}}$. From the MSD, we can either infer the full XXX (Mathias) model, based on XXX equation:

$$4\sigma^2 + R_\infty^2\left(1 - \exp\left(-\frac{4dx}{R_\infty^2}\right)\right) \tag{7}$$

or simply approximate the `DCon2_WT1` (XXX) with half of the slope of the first three data points of the MSD (Mathias, why half?).

We can compute the energy, U, in two different ways; U_{left} and U_{right}. The first method is based on a geometric calculation depending on the fraction of the slow diffusion coefficient from the Wild

Type 1 calculation, $\theta_{\text{slow}}^{\text{WT}_1} \equiv \theta_1^{\text{WT}_1}$:

$$
\begin{aligned}
V_{\text{cap}} &= \frac{\pi h^2}{3(3r_0 - h)} \\
V_0 &= \frac{4\pi}{3 - 2V_{\text{cap}}} \\
V_F &= \frac{8V_0}{4\pi/3} \frac{4\pi}{3R_R^3} \\
U_{\text{left}} &= -\log\left(\theta_{\text{slow}} \frac{V_0 - V_F}{(1 - \theta_{\text{slow}}^{\text{WT}_1})V_F}\right),
\end{aligned}
\tag{8}
$$

where $r_0 = 1.0$, $h = 0.85$, and $R_R = 0.13$.

The other energy, U_{right}, can be calculated from the value of `DCon2_WT1` (half slope) from Wild Type 1, the `Db_focus` (half slope) from the Focus files, and the fast diffusion coefficient from the delta files: $\theta_{\text{fast}}^{\text{delta}} \equiv \theta_2^{\text{delta}}$:

$$
U_{\text{left}} = \log\left(\frac{\texttt{DCon2\_WT1} - \texttt{Db\_focus}}{\theta_{\text{fast}}^{\text{delta}} - \texttt{Db\_focus}}\right). \tag{9}
$$

2.4 | Implementation

The data analysis has been carried out in Julia (Bezanson et al., 2017) and the Bayesian models are computed using the Turing.jl package (Ge, Xu, and Ghahramani, 2018). We use Hamiltonian Monte Carlo sampling (Betancourt, 2018) with the NUTS algorithm (Hoffman and Gelman, 2011). In particular, each Bayesian model have been run with 4 chains, each chain 1000 iterations long after discarding the initial 1000 samples ("warm up").

3 | RESULTS

3.1 | Dynamics of SIR3 reveals two dominating populations of the motion

We started out by investigating the mobility of individual SIR3 molecules in vivo. Here we typically have 5-8 repair foci. To image SIR3 without changing its normal level, we generated haploid cells expressing the endogenous SIR3 fused to Halo (Figure 1A). Before we wanted to visualize the cells on a PALM microscope (see Materials and methods), we incubated the exponentially growing cells with fluorescent and fluorogenic JF647, a dye emitting light only once bound to SIR3. We were very used a low concentration of JF646 allowing for the observation of individual molecules (Ranjan et al., 2020; Figure 1— figure supplement 2). Rad52-Halo bound to JF646 (Rad52-Halo/JF646) were visualized at 20 ms time intervals (50 Hz) in 2-dimensions during 1000 frames until no signal was visible.

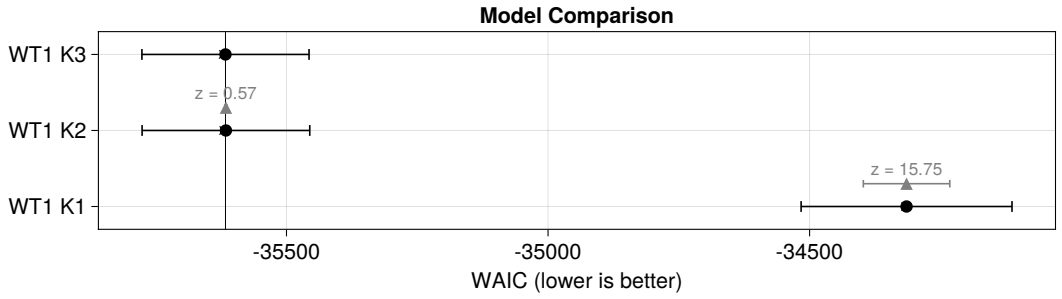


Figure 2. Comparison between diffusion models with $K = 1$, $K = 2$, or $K = 3$ diffusion coefficients for the Wild Type 1 data (WT1). The x-axis shows the WAIC score, where lower values indicate higher-performing models. The WAIC-score for each model is shown in black along with its uncertainty. The difference in WAIC-scores between the model and the best performing model (WT1 K3) is shown in grey with z being the number of standard deviations between them.

3.2 | Bayesian Analysis

Comparing the three diffusion models with 1, 2, or 3 diffusion coefficients, respectively, we find that the model with only a single diffusion component is simply not advanced enough to fully explain the data, see Figure 2. This figure shows that, even though the 3 component model is the best-performing of the models, when judging by the number of standard deviations, z, that the best model's WAIC is higher than the second best model's WAIC, it is statistically non-significant ($z < 2$). Since the performance of both the 2 and 3 component models are indistinguishable, we follow Occam's razor and continue with the former model.

Bayesian models allow for far greater flexibility than traditional frequentist models, including internal validation checks and diagnostic criteria to make sure that the model has not converged. In particular, we made sure that all $\hat{R}$-values were less than 1.01. To fully validate the $K = 2$ model, we show the traceplots and posterior distributions for the different parameters in Figure 3. The left part of the figure shows the parameter estimate as a function of MCMC iteration, i.e. traceplot, which, for correctly sampled chains, should resemble a fuzzy caterpillar (and not a skyline which would indicate bad mixing) (Roy, 2020). We find that the slow diffusion coefficient for WT1 data is: $\theta_{\text{slow}}^{\text{WT}_1} = 0.0417 \pm 0.0014\,\text{XXXunit}$.

XXX Energy computation shows that, see Figure 4.

Final results:

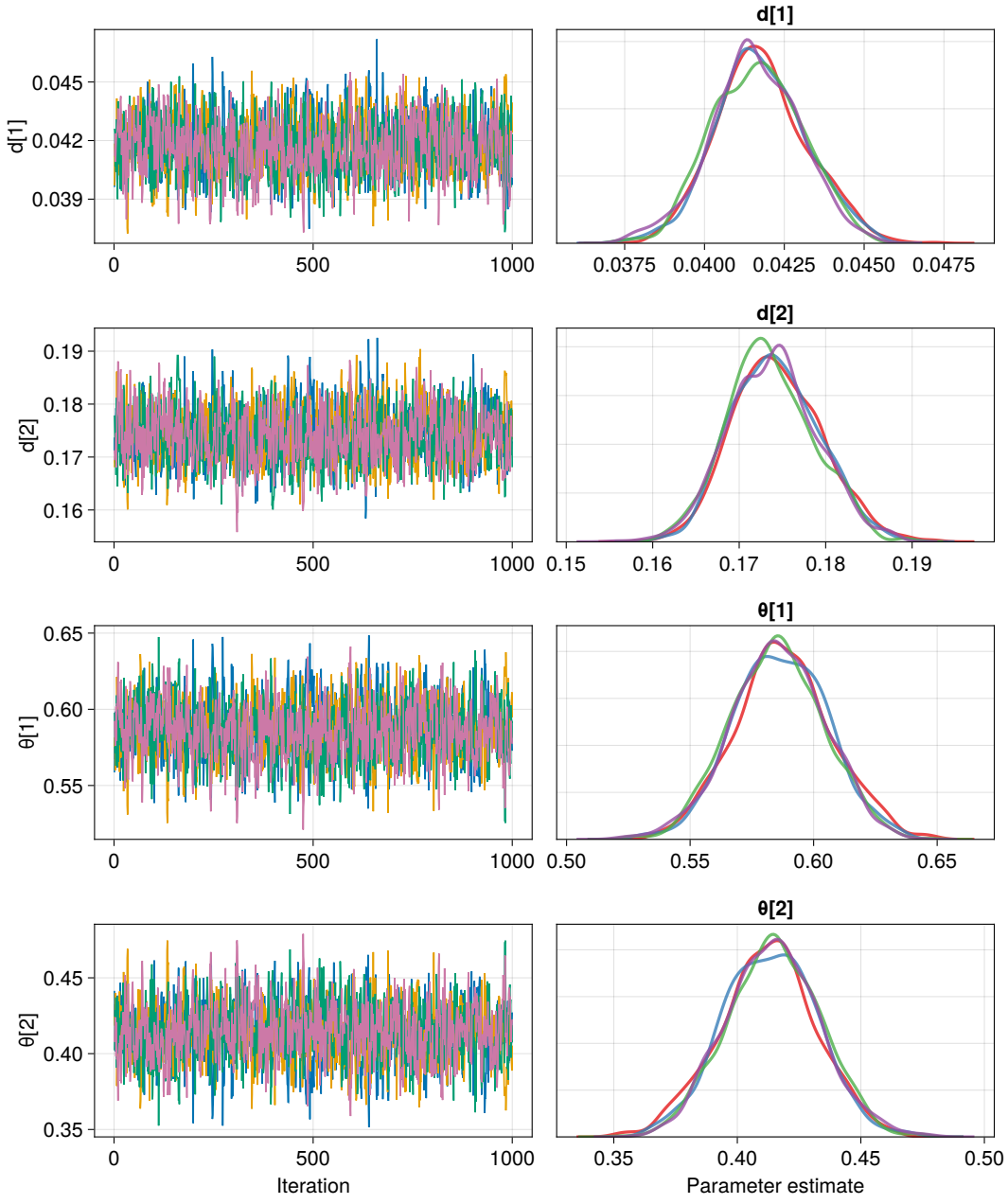


Figure 3. Results of the $K = 2$ diffusion model. Left) Traceplots. Right) Density plots.

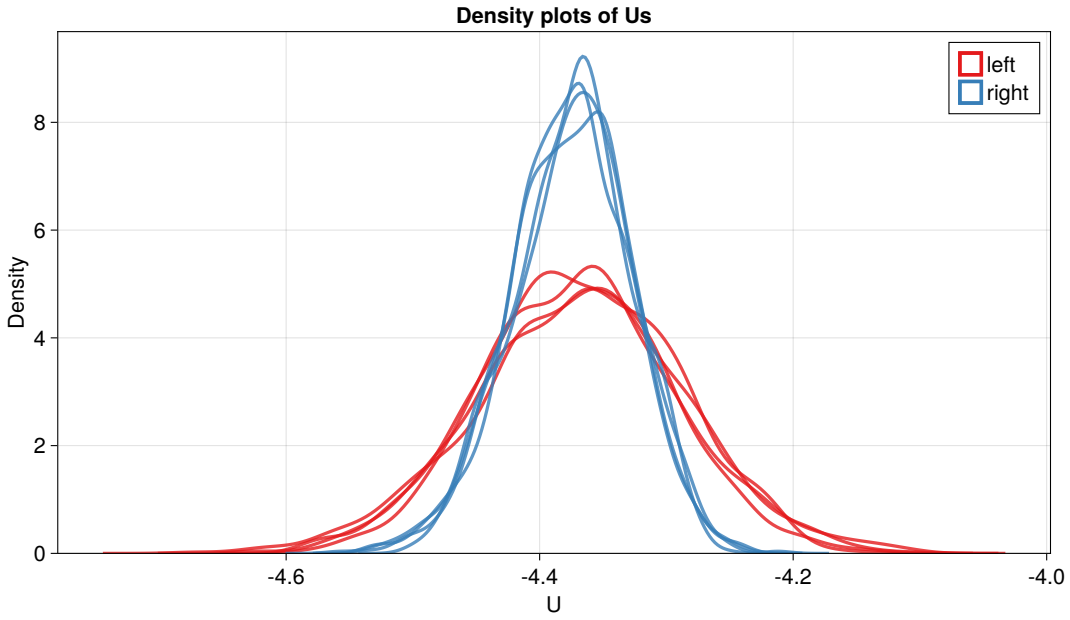


Figure 4. Density plot of the energy, U, using either the left or the right computation approach (XXX Mathias). The energy computed using the left computation is shown in red, and in blue with the right computation. The four different MCMC chains (for each approach) are shown as individual lines.

$$
\begin{aligned}
\texttt{DCon1\_WT1} &= 0.010\,29 \pm 0.001\,83 \\
\texttt{DCon2\_WT1} &= 0.017\,334\,3 \pm 0.000\,000\,4 \\
\texttt{Db\_focus} &= 0.006\,629\,0 \pm 0.000\,000\,1 \\
U_{\text{left}} &= -4.373 \pm 0.078 \\
U_{\text{right}} &= -4.375 \pm 0.047 \\
\texttt{Din\_hyper\_WT} &= 0.031\,51 \pm 0.002\,71 \\
\texttt{DCon1\_hyper\_WT} &= 0.008\,14 \pm 0.002\,34 \\
\texttt{DCon2\_hyper\_WT} &= 0.015\,391\,1 \pm 0.000\,001\,8 \\
\texttt{Db\_hyper\_focus} &= 0.001\,315\,7 \\
U_{\text{left-hyper}} &= -4.364 \pm 0.204 \\
U_{\text{right-hyper}} &= -4.108 \pm 0.047
\end{aligned}
\tag{10}
$$

4 | DISCUSSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

4.1 | Acknowledgment

Acknowledgements here

4.2 | Data availability

Source code is hosted at GitHub: https://github.com/ChristianMichelsen/diffusion.

REFERENCES

Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1. Publisher: SIAM, pp. 65–98. URL: https://julialang.org/.

Dietz, Laura (2022). "Directed factor graph notation for generative models". In.

Ge, Hong, Kai Xu, and Zoubin Ghahramani (2018). "Turing: A Language for Flexible Probabilistic Inference". en. In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, pp. 1682–1690. URL: https://proceedings.mlr.press/v84/ge18b.html (visited on 2022).

Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). "Understanding predictive information criteria for Bayesian models". en. In: *Statistics and Computing* 24.6, pp. 997–1016. ISSN: 1573-1375. DOI: 10.1007/s11222-013-9416-2. URL: https://doi.org/10.1007/s11222-013-9416-2 (visited on 2022).

Hoffman, Matthew D. and Andrew Gelman (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *arXiv:1111.4246 [cs, stat]*. arXiv: 1111.4246.

McElreath, Richard (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.

McLachlan, Geoffrey J. and David Peel (2004). *Finite Mixture Models*. en. Google-Books-ID: c2_fAox0DQoC. John Wiley & Sons. ISBN: 978-0-471-65406-3.

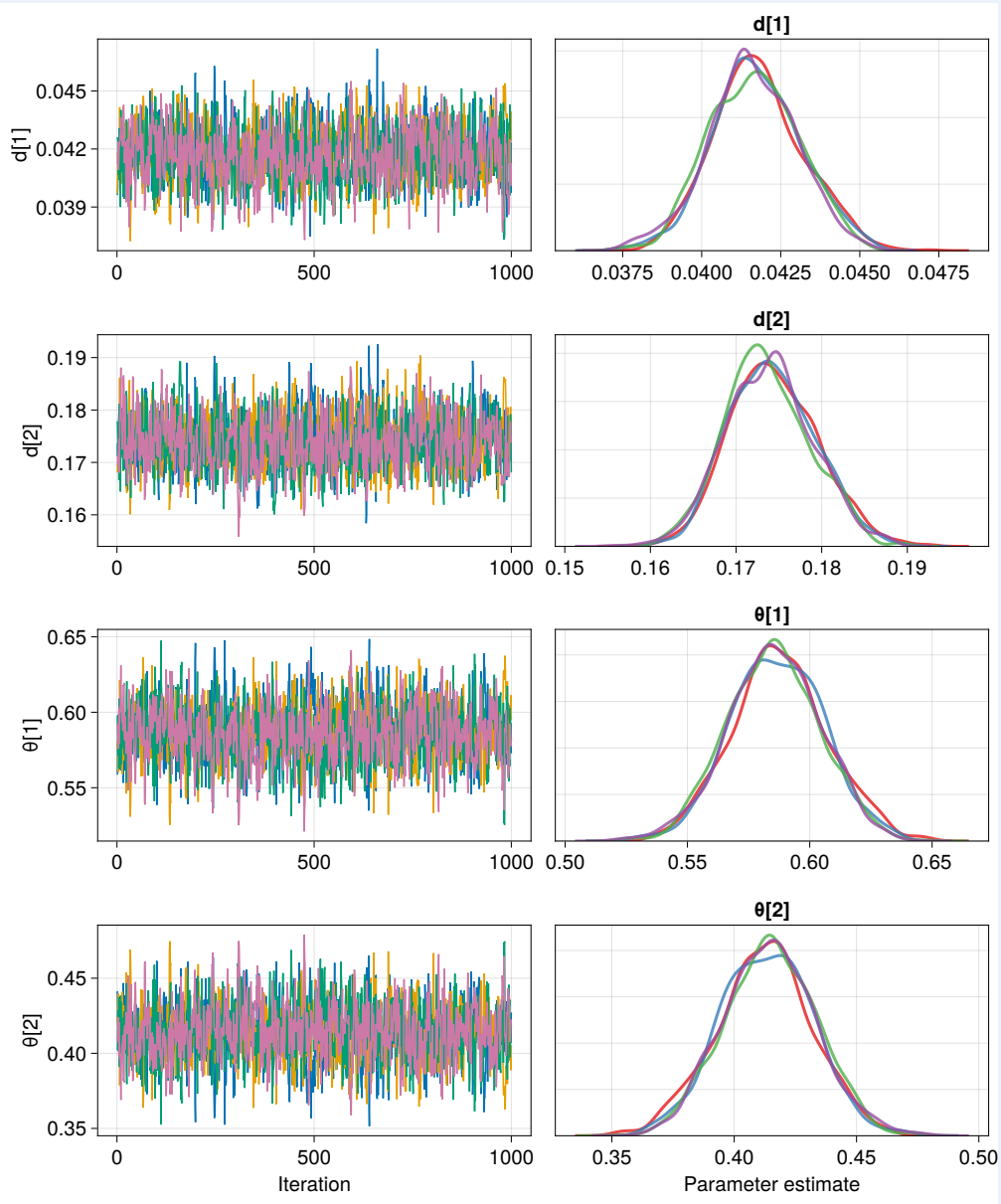
Roy, Vivekananda (2020). "Convergence Diagnostics for Markov Chain Monte Carlo". In: *Annual Review of Statistics and Its Application* 7.1. _eprint: https://doi.org/10.1146/annurev-statistics-031219-041300, pp. 387–412. DOI: 10.1146/annurev-statistics-031219-041300. URL: https://doi.org/10.1146/annurev-statistics-031219-041300 (visited on 2022).

Watanabe, Sumio (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *Journal of Machine Learning Research* 11.116, pp. 3571–3594. ISSN: 1533-7928.

Appendix 1

A | APPENDIX FIGURE 1

Here an example of an appendix figure.



Appendix 1—figure 1. Blablabla.