Bayesian Probabilistic Analysis of DEER Spectroscopy Data
Stephan Pribitzer, Sarah Sweger, Stefan Stoll
Department of Chemistry, University of Washington, Seattle

Single Gaussian analysis

The probabilistic analysis method is first shown for a synthetic data set with low complexity. We used a synthetic distance probability distribution that is taken from the large simulated 1G trace (T1L) test data [1], which resembles a single Gaussian.

From this PMF we generated two noisy signal traces, one with favorable values and another with less ideal values for the modulus depth, background decay rate, trace length and noise level.

The figure on the right shows the results of the Bayesian analysis for both cases. The top row depicts the marginalized posterior for each parameter for the poor case. The first column shows the same for the good case. Dashed lines show the true parameter values.

Both parameter posteriors show the outcome of the Bayesian analysis most directly, the column shows the same for the good case. Dashed lines represent the ground truths.

The recorded time domain trace exhibits modulations that depend on the dipolar couplings that are present. Similar to a Fourier transform, these oscillations can be transformed into the distance domain, yielding a distance PMF of the interpoint distance. This step constitutes an inverse problem that is usually solved with Tikhonov regularization.

DEER theory

The noise-free DEER signal is

\[ \tilde{V}(\mathbf{r}) = V_{\text{true}}(\mathbf{r}) \exp(-i \mathbf{K}_1 \mathbf{r}) \]

where \( \mathbf{r} \) is the position of the pulse pair and \( V_{\text{true}} \) is the echo amplitude in absence of the pump pulse. \( V_{\text{true}}(\mathbf{r}) \) is the intramolecular modulation function given by

\[ V_{\text{true}}(\mathbf{r}) = \exp(-i \mathbf{K}_1 \mathbf{r}) \]

with the modulation depth \( K_1 \) and the normalized distribution \( P(\mathbf{r}) \) of the spin-spin distance \( r \).

\[ P(\mathbf{r}) = \frac{1}{\sigma_\mathbf{r}} \exp\left(-\frac{r^2}{2 \sigma_\mathbf{r}^2}\right) \]

where \( \sigma_\mathbf{r} \) is the distance decay rate constant.

One way to approximate the spin distance distribution \( P(\mathbf{r}) \) is a linear combination of normalized Gaussian basis functions:

\[ P(\mathbf{r}) = \sum_i a_i \phi_i(\mathbf{r}) \]

where \( a_i \) is the number of Gaussians, \( \phi_i \) the centers of the Gaussians, and \( \sigma_i \) are the full widths at half maximum.

In vector form, each of these data points can be written as a random sample from a Gaussian distribution with center \( \mu_i \) and covariance matrix \( \Sigma_i \), where \( \mu_i \) is the noise level:

\[ V = \mu_i + \mathbf{N} \]

The ultimate goal of DEER spectroscopy is to obtain a probability distribution of the spin-spin distance, ideally including some sort of uncertainty assessment.

Synthetic data

The main tool for spin distance distributions encountered in DEER spectroscopy of proteins are asymmetric and multimodal, and therefore poorly approximated by a single Gaussian. We analyzed a noisy time trace generated from a bimodal distribution from the T4L test set, with two distinct modes, one significantly weaker than the other.

The example shown here is a challenging, broad distribution with several poorly resolved modes of similar intensity. The Bayesian analysis was conducted using three Gaussian distribution models.

The pair of marginalized posteriors present an overwhelming amount of information. To help with interpretation it is advantageous to condense them to matrices of pairwise Pearson correlation coefficients.

Multimodal distributions and model selection

Both the 1G and 2G model show systematic deviations, even though they provide an apparently precise PMF. The time domain signal is well-described by both the 1G and 2G models. Though not ideal, yet useful in this situation, a formal approach for model comparison and to identify overfitting is the calculation of the Bayes factors [3]. The Bayes factor is the ratio of the posterior odds of two models:

\[ \text{Bayes factor} = \frac{P(Y|M_1)}{P(Y|M_2)} \]

A log Bayes factor of 0 can be seen as a relatively strong indication for \( M_1 \) over \( M_2 \). The Bayes factors \( \log_{10} P(Y|M) \) for all 1G–2G models relative to the 1G model, give preference for 3G and 4G. However, the relatively small Bayes factor \( \log_{10} P(Y|M) \approx 1 \) indicates that the 4G is not really preferred over the data.

Both models and Bayes factors are required for complete analysis. While the former diagnostics systematically project the latter can identify overfitting.

Acknowledgments

This work was supported by the National Science Foundation with Grant CHE-1452087 (S.L.) and the National Institutes of Health with Grants RO1 GM120825 and RO1 GM123099 (S.L.).


Markov chain Monte Carlo (MCMC) sampling

Drawing from the posterior we are interested in the posterior probability distribution of the parameter vector \( \theta \) conditioned on the measured signal \( y \), model \( M \) and any additional information. The posterior can be expressed as

\[ p(\theta|M, y) = p(y|\theta,M) p(\theta|M) \]

Here, \( p(\theta|M, y) \) is the likelihood

\[ p(y|\theta,M) = \prod_i p(y_i|\theta,M) \]

which quantifies the degree of fit between the data and the model.

For each parameter vector \( \theta \) we can write prior probability distributions that can be combined

\[ p(y|M, \theta, y) \]

For the Markov chain Monte Carlo (MCMC) sampling from the posterior we use pymc3 with a NUTS sampler. Each chain is initialized with different starting points and propagated for 5 000 steps to tune the sampler. The chains are then propagated for 20 000 – 80 000 steps. Convergence is assessed using the rank normal split-\( R \) statistic [2].

Prior distributions

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 )</td>
<td>Unif(0, 100)</td>
</tr>
<tr>
<td>( r_2 )</td>
<td>Unif(0, 100)</td>
</tr>
<tr>
<td>( r_3 )</td>
<td>Unif(0, 100)</td>
</tr>
</tbody>
</table>

Posterior distributions

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 )</td>
<td>Unif(0, 100)</td>
</tr>
<tr>
<td>( r_2 )</td>
<td>Unif(0, 100)</td>
</tr>
<tr>
<td>( r_3 )</td>
<td>Unif(0, 100)</td>
</tr>
</tbody>
</table>

Further reading