Challenges and opportunities for Bayesian statistics in proteomics

Oliver M. Crook * [1], Chun-wa Chung[2], and Charlotte M. Deane[1]

[1] *Department of Statistics, University of Oxford, Oxford, UK*
[2] *Structural and Biophysical Sciences, GlaxoSmithKline R&D, Stevenage, UK*

October 18, 2021

Abstract

Proteomics is a data-rich science with complex experimental designs and an intricate measurement process. To obtain insights from the large datasets produced, statistical methods, including machine learning are routinely applied. For a quantity of interest, many of these approaches only produce a point estimate, such as a mean, leaving little room for more nuanced interpretations. By contrast, Bayesian statistics allows quantification of uncertainty through the use of probability distributions. These probability distributions enable scientists to ask complex questions of their proteomics data. Bayesian statistics also offers a modular framework for data analysis by making dependencies between data and parameters explicit. Hence, specifying complex hierarchies of parameter dependencies is straightforward in the Bayesian framework. This allows us to use statistical methodology which equals, rather than neglects, the sophistication of experimental design and instrumentation present in proteomics. Here, we review Bayesian methods applied to proteomics, demonstrating its potential power, alongside the challenges posed by adopting this new statistical framework. To illustrate our review, we give a walk-through of the development of a Bayesian model for dynamic organic orthogonal phase-separation (OOPS) data.

1 Introduction

Decision making spans the entire research process. Ultimately, it is a choice to believe an explanation for a phenomena given the current evidence. For some theories, the evidence is overwhelming: careful mechanistic experiments and verifiable model predictions have never contradicted that theory[1]. This scenario is, however, rare. In practice, we make decisions under uncertainty and the evidence is not clear-cut[2]. Bayesian statistics allows us to make

*oliver.crook@stats.ox.ac.uk

inferences from that evidence to enable decision making in those cases[3]. In contrast to *frequentist* methods, Bayesian inference allows us to use probability to model degrees of belief, rather than just frequencies. Consequently, models that are consistent with the available evidence are more probable and incompatible models are less probable[4,5]. By using probability theory in this manner, there is a recipe for taking *prior beliefs* (i.e. information encoded by domain expertise) and updating them to *posterior beliefs* using observed data[3]. This posterior probability distribution quantifies the models compatible with domain expertise and our experimental data[4]. This recipe is known more formally as *Bayes' theorem*.

Mass-spectrometry-based proteomics is a complex scientific field[6]. The techniques versatility allows it to explore differential abundance[6], protein turnover[7], interactions[8], thermal stability[9], structure[10,11], spatial information[12,13,14] and more[15,16,17]. In each case, data are manipulated, thresholded and filtered so that a statistical test or machine learning algorithm can be applied. The results are then frequently a single value, which is granted the role of arbiter of truth. Bayesian statistics allows the propagation or quantification of uncertainty in all of the step of an analysis, replacing the current implicit ad-hoc approaches with explicit models and summarises the output with a probability distribution consistent with the data [4]. This paradigm progression not only provides an ability to ask new questions of the data but a consistent way to perform inference and criticize models[3,4].

Bayesian statistics offers considerable potential for the examination of proteomics data; despite this, it has not been readily adopted in the community. Here, we review the contribution Bayesian statistics has already made to proteomics, clarify the Bayesian workflow and how it can be applied, highlight a number of modelling strategies and outline current challenges for the proteomics community. Throughout, we illustrate our analysis with examples from the proteomics literature, focusing on building a model for dynamic organic orthogonal phase separation (oops) data[18].

2 Main

2.1 Bayes, in brief

Before reviewing the contributions Bayesian statistics has already made to proteomics, we introduce the technical background and notation. We use $P(E)$ to denote the probability of the event E. E can be anything from "it rains tomorrow" to "my parameter falls between the values a and b". We let D be notation for the observed data, for example from a shotgun proteomics experiment and let x be a data point from D, such as a measurement for a particular protein. We assume that x is a sample from some probability distribution p and we write $x \sim p(x|\theta)$, for example this could be a log normal distribution. Let α be hyperparameters of the parameter distribution, such that the θ themselves are drawn from a probability distribution $\theta \sim p(\theta|\alpha)$. For example, the mean of the log normal distribution could be drawn from a normal distribution.

The *prior distribution* captures our domain expertise and is the distribution of the parameters before any data is observed: $p(\theta|\alpha)$. The *prior* could capture, say, that abundance values are positive and are unlikely to exceed the number of grains of sand on Earth. The sampling distribution is the distribution of the data given the parameters $p(D|\theta)$, we can write this as a function $L(\theta|D)$ called the *likelihood*. If we average (or *marginalise*) the distribution of the data over the parameters, we obtain the so-called marginal likelihood:

$$p(D|\alpha) = \int L(\theta|D)p(\theta|\alpha)\,\mathrm{d}\theta. \tag{1}$$

The posterior distribution of the parameters is determined by Bayes' theorem, as the following:

$$p(\theta|D,\alpha) = \frac{p(D|\theta,\alpha)p(\theta|\alpha)}{p(D|\alpha)}. \tag{2}$$

Bayes' theorem tells us the mathematical way to update beliefs in light of evidence: simply multiply our prior and likelihood and renormalise by the marginal likelihood. Bayes' theorem implies a self-consistency property: the posterior averaged over the data returns the prior[19]:

$$p(\theta) = \int\int p(\theta|\tilde{y})p(\tilde{y}|\tilde{\theta})p(\tilde{\theta})\,\mathrm{d}\tilde{y}\mathrm{d}\tilde{\theta}, \tag{3}$$

where $\tilde{y} \sim p(D|\tilde{\theta})$. In Bayesian analysis to perform prediction, instead of simply taking a single parameter value forward, we use averaging:

$$p(\tilde{x}|D,\alpha) = \int p(\tilde{x}|\theta)p(\theta|X,\alpha)\,\mathrm{d}\theta. \tag{4}$$

In summary, Bayesian statistics provides us with a distribution of plausible parameter values from the *posterior* and a distribution of hypothetical predicted values from the *posterior predictive distribution*. We can then ask bespoke question of these probability distributions; for example, $P(\theta > 2|D,\alpha) = \int_2^\infty p(\theta|D\,\alpha)\,\mathrm{d}\theta$ is the probability that a parameter is greater than 2. For proteomics, these could be the probability that a fold-change exceeded a certain value or the probability that a spectra belongs to a particular peptide.

2.2 Bayesian contributions to proteomics

2.2.1 Bottom-up proteomics and differential abundance

Next, we discuss Bayesian approaches already applied to proteomics. We focus on proteomics data generated via mass-spectrometry but refer to contributions for the reverse-phase-protein-array (RPPA) literature[20,21,22] and 2D gel electrophoresis[23]. A number of approaches have been aimed at quantification and differential abundance analysis of bottom-up proteomics data[24,25,26,27,28,29,30,31,32,33]. Carvalho et al. [33] describe the use of Bayesian statistics to improve analysis of spectral counting data using a Poisson likelihood and calculating the probability of detecting a protein in a particular sample. However, they do not

exploit the full Bayesian toolkit and simply interpret their probabilities as p-values. Thus, it is not clear whether this is any better than simply using a likelihood-based method. A number of approaches[25,26,28,32] argue for propagating and quantifying the uncertainty in the analysis rather than simply the underlying quantitation values, either through relaxing the parsimony assumption[32], including ion statistics via a two-level Beta-Binomial model[28], or jointly modelling identification and quantitation statistics[25,26]. Millikin et al. [29] use an analogue of the t-test and exploit the posterior distribution by using an interval thresholding approach. O'Brien et al. [31] note the inherent compositional nature of labelled proteomics approaches and include a modelling parameter to model ratio compression allowing better estimation of the true fold changes. Importantly this parameter is shared across all proteins, which allows them to estimate the parameter accurately even for proteins with few observed peptides. Jow et al. [34] also model isobaric laballed mass-spectrometry data but do not model ratio compression. Meanwhile for label-free experiments, O'Brien et al. [35] explicitly model missingness showing that jointly modelling missingness and abundance leads to improved performance. All of these approaches demonstrate some benefit over previously applied methods suggesting that combining these methods would provide further improvements. It also suggests translating these methods to other proteomics techniques would be a fruitful endeavour.

2.2.2 Protein and peptide identification

One of the fundamental problems in mass-spectrometry based proteomics is identifying a peptide from a spectra. A spectra can be very noisy and $b-, y-$ ions can be missing which results in a complex observation process. Furthermore, we have prior knowledge of observing particular amino acid sequences and knowledge of the cleavage process. This is an ideal scenario for the application of Bayesian methods. Indeed, a number of approaches have been applied[36,37,38,39]. Chen et al. [36] used a fairly simple framework to calculate peptide identification probabilities based on peptide coordinates. Halloran et al. [37] employed a dynamic Bayesian network in a method called DRIP, allowing for insertions and deletions to the spectra. By modelling possible alignments between theoretical and observed spectra they were able to calculate the most probable peptide match. The authors found that their approach was an improvement over available methods particularly for low resolution MS2 data. Lewis et al. [38] took a different approach to the same problem, incorporating a scoring function into a likelihood model. Their model also allowed deletions directly via indicator functions. Insertions were characterised by excessive deviations from the spectra of the candidate peptide, where excessive is characterised probabilistically via laplace noise. The authors also included prior information about possible cleavage pairs, as well as prior information about the probability of observing a particular peptide sequence in the dataset. Finally, in constrast to Halloran et al. [37] they made full use of Bayesian methods and provide a posterior distribution over possible peptides and parameters. This could have allowed multiple peptides to be associated to a spectrum with differing certainty which

could have been used in downstream analysis. Claassen et al. [39] tackled a slightly different problem and used a non-parametric Bayesian model to predict the coverage in sequential LC-MS/MS experiments but suggest their approach could also have been adapted to database searching and de novo sequencing. Again, these approaches are all shown to have benefits over previously applied methodology. The clearest is the inclusion of more information and the ability to provide a flexible, and well rationalised, model to the underlying data. The ability to exploit uncertainty captured by the posterior distribution for downstream analysis is far more insightful than simply point estimates from a Bayesian analysis. However, these approaches have not yet been widely adopted by the community. This could be because the methods are difficult or expensive to apply, or the benefits are not compelling. We aim to show throughout this review benefits of a Bayesian analysis are compelling and straightforward to obtain.

2.2.3 Proteoforms and post-translation modifications

A number of approaches are interested in applications to proteoform analysis (splice isoforms) or post-translational modifications[40,41,42,43,44]. Chung et al. [40] employed a non-parametric mixture model to jointly model the modification mass for each PTM group and the true (unobserved) location of the modified amino acid. Their approach outperformed other approaches, is fully automated and provides modification confidence scores. However, the approach did not model the underlying spectrum, which could have resulted in unnecessary false positives. Webb-Robertson et al. [41] tackled the proteoform problem by deconvolving peptides into signatures even if they are associated with the same protein. However, they only used a Bayesian point estimate rather than exploiting the full posterior distribution. Lim et al. [42] used a Bayesian model to estimate the phosphorylation stoichiometry using Bayesian statistics. By incorporating a physically plausible model they removed problems with previous models that could have allowed negative stoichiometry. Their joint model allowed them to borrow power across replicates and they reported downstream uncertainty. Shteynberg et al. [43] use a Bayesian mixture model to compute probabilities for modification sites. This allowed them to combine precomputed scores in a rational way but, again, they did not examine the full posterior distributions. Mallikarjun et al. [44] employed a Bayesian linear regression modelling strategy to analyse differential PTM data, suggesting their approach outperformed other methods and could allow uncertainty in missing values. The main benefit here appeared to be the regularisation of the parameters using priors rather than specifically the uncertainty quantification in the analysis.

2.2.4 Biomarkers and clinical proteomics

Protein biomarkers, molecular indicators of aberrant processes or disease, and clinical proteomics are a key component of proteomics research. For a review of Bayesian method development in biomarker discovery, see Hernández et al. [45]. Morris et al. [46, 47] developed Bayesian wavelet-based functional mixed models for mass-spectrometry-based proteomics

data. Their advanced framework, allowed the simultaneous use of nonparametric fixed and random effects, which facilitated adjustment for clinical and experimental covariates that could affect the intensity and location of a spectra. Working with posterior distributions they were able to compute important quantities such as the probability of intensity changes for fixed fold levels and were able to control a Bayesian false discovery rate; that is, the posterior probabilities are thresholded to control the error rate. Liao et al. [48] combined that framework with image analysis methods to enable biomarker discovery from LC-MS data. Hwang et al. [49] developed a pipeline, MS-BID, for biomarker analysis which uses a Bayesian analysis of variance (ANOVA). Harris et al. [50] applied a Bayesian hierarchical linear probit regression (regression where only two outcomes are allowed) model to determine discriminative biomarkers from mass spectrometry data. They found their approach improved over a simple K-nearest neighbour method. Furthermore, by using posterior probabilities they were able to determine which samples will be the most promising for prognostics. Kuschner et al. [51] demonstrated a Bayesian network to perform feature selection from mass-spectrometry data. The selected features then provided excellent predictive power. Though again this approach still used a Bayesian point estimate and instead of full posterior distributions. Deng et al. [52] developed a Bayesian network which allows them to integrate mass-spectrometry and microarray data, allowing them to borrow power between mRNA and protein levels. Here, they made use of the flexibility of Bayesian statistics to incorporate different modalities and weigh up the uncertainty between different datasets. More recently Liu et al. [53] developed Bayesian Function-on-Scalar quantile regression for mass-spectrometry data. This approach noted that biomarker difference may not be apparent at mean regression but rather at a particular quantile (such as the 0.95 quantile). It simultaneously accounted for the functional nature of MALDI-TOF data, incorporated prior knowledge for adaptive regularization and a basis representation which allowed borrowing of power. They found their method identifies biomarkers overlooked by mean regression.

Bayesian methods for biomarker and clinical proteomics are more developed than other examined proteomics sub-fields with several exemplary methods that make full use of the flexibility of Bayesian modelling and the rich output of the posterior distribution.

2.2.5 Chromatography

To facilitate peptide identification in mass-spectrometry a liquid-chromatography step is usually applied. The time at which a peptide elutes from the liquid chromatography, called the retention time, can be used as additional information to help identify peptides. However, there is uncertainty in this retention time and they can vary from one run to another. Chen et al. [54] developed a Bayesian model called DART-ID, which models a latent (unobserved) global retention time alignment. This alignment allowed them to combine the outputted posterior error probability of MaxQuant with the inferred RT density in each experiment. Hence, by using this result they updated their confidences and improved coverage in experiments by 50%. Whilst their approach is powerful, they only used a point estimate and

obtained uncertainty through bootstrapping. Maboudi Afkham et al. [55] are interested in the uncertainty in peptide retention time measurements. Using a Gaussian process regression method they were able to accurately predict retention times and obtain uncertainty estimates. They then used the posterior distribution from the regression analysis as a variable retention time window to identify potentially incorrect peptides. This improved over fixed windowing strategies. One potential strategy for improvement, in a similar vain to DART-ID, would have been to update the identification probabilities based the deviation probability from the predicted retention time. This approach naturally fits within a Bayesian framework.

2.2.6 Intact, top-down and structural proteomics

Saltzberg et al. [56] proposed a Bayesian model to resolve residue level information from hydrogen-deuterium exchange mass-spectrometry. They chose uninformative priors, and though they performed inference using Monte-Carlo methods they did not use the posterior distribution. Furthermore, they do not justify why there model allows for negative deuterium incorporation, which maybe arises from a misunderstanding of the positivity constraint induced by their exponential likelihood model. Proteoform analysis is one of the key challenges in top-down proteomics. LeDuc et al. [57] introduced a C-score, not be confused with a C-statistic, to facilitate automated identification and characeterisation of proteoforms from top-down proteomics data. Ultimately, their approach allowed them to rank probable proteoforms having observed their data. Performing an analysis in a Bayesian framework allowed them to specify a generative model, provide expert prior information and carefully model the underlying noise distribution. Their proposed C-score is essentially a transformed posterior error probability. However, despite their Bayesian framework, they opted for a point estimate of their model, which could have been greatly enhanced by examining the full posterior of their model. Marty et al. [58] proposed a Bayesian deconvolution algorithm for Ion Mobility spectra, which was extended in Kostelic et al. [59]. Their approach allowed the convolution of the charge distribution with the peak shape to obtain a flexible deconvolution approach. The wide extent of their applications demonstrated the clear benefits of their method. However, their approach also used a point estimate from their analysis. Hence, apart from the use of prior information, it is not clear what particular benefit a Bayesian analysis had for their approach.

2.2.7 Functional proteomics

Functional proteomics methods aim to decipher protein-function on a system-wide scale. One approach is spatial subcellular proteomics[13,60] where protein are localised to their subcellular niche using mass-spectrometry data. Bayesian approaches have been developed for biochemical fractionation-based subcellular proteomics[20,61,62,63,64]. Crook et al. [20, 61, 62] demonstrated Bayesian modelling can quantify uncertainty in protein subcellular localisation and identify cases where this may correspond to multi-localising proteins. Crook et al. [61]

showed that a even a Bayesian point estimate may overlook these cases and more information is obtained by examining the full posterior distribution. Crook et al. [63] allowed the uncertainty in the number of subcellular niches to be accounted for and showed that allowing additional niches can be uncovered. However, the model appeared sensitive to the prior choices and should be chosen carefully. Crook et al. [64] built on these experiments to analyse differential localisation experiments showing that modelling uncertainty improved power and interpretation compared with other methods. This fully Bayesian analysis; however, is computationally intensive as it attempts to model many datasets at once. Another functional approach is affinity purification mass spectrometry (AP-MS), which allows us to determine protein interactions and complexes[60]. Choi et al. [65] developed a non-parametric Bayesian model to bi-cluster AP-MS data. They sampled from the posterior distribution and are hence able to report the uncertainty in the clustering. However, their nested model assumed that the conditional on the Bait cluster the Prey clusters are independent and their model assumed exchangability (permutation leads to the same probability distribution) of the rows and columns. Fang et al. [66] proposed a semi-parametric model for thermal protein profiling after identifying proteins that deviate from classic sigmoid behaviour. Semi-parametric models combine interpretable parametric models with more flexible non-parametric models. Using Bayesian analysis they critically assessed the semi-parametric and parametric model fits and demonstrate those proteins that are better modelled by the semi-parametric model share functional enrichments. Again this fully Bayesian approach had demanding computational requirements, which may explain why many methods choose not to employ Bayesian methods.

2.3 The Bayesian workflow

2.3.1 Motivating example

To illustrate the Bayesian workflow, we examine some recently introduced proteomics data generated using the orthogonal organic phase separation (oops) method of Queiroz et al. [18]. This method is able to efficiently enrich for RNA-binding proteins and hence, by adapting to the dynamic setting, is able to examine differential RNA binding. This is where the proportion of a particular protein bound to RNA changes depending on the condition. Here, we examine an experiment where thymidine-nocodazole was used to induce cellular arrest. Total and oops-enriched protein abundances where obtained at $0, 6$ and 23 hours post treatment. Each experiment was performed in triplicate, except for at 6h when four replicates were taken. The 10 total and 10 oops samples were labelled using 2 separate TMT 10-plex kits and quantitative mass spectrometry was performed in two runs using SPS-MS3 on an Orbitrap Fusion Lumos. Here, we attempt to use the Bayesian toolkit to model this data and answer questions about changes in RNA-binding. A heatmap of the data is shown in figure 3. A protein was chosen at random to illustrate the modelling process, NCAPD2, a regulatory subunit of the condensin complex. NCAPD2 is known to have differential

subcellular localisation throughout the cell-cycle[67].

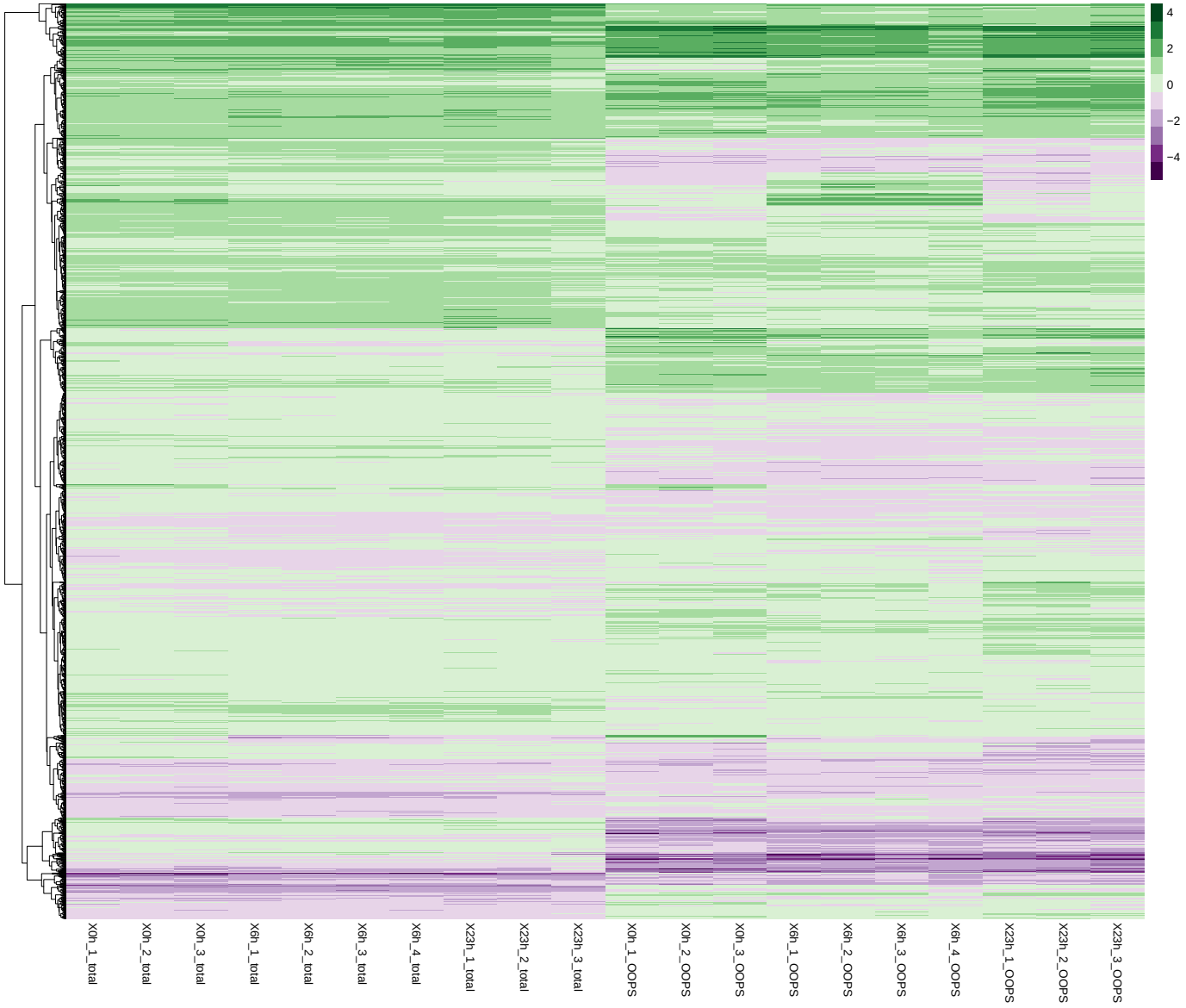

Figure 1: **Exploratory data analysis of OOPS.** A heatmap of the mass-spectrometry data generated by the oops experiment. The tree clustering is produced using Ward's method. Each cell represents z-score normalised protein abundances. Column annotations are encoded as Xtime_replicate number_sample type

.

2.3.2 Generative modelling

Having highlighted the successes and limitations of some of the contributions of Bayesian methods to mass spectrometry-based proteomics, below we outline the Bayesian workflow

to facilitate it for proteomics. The first tension of Bayesian analysis is the pairing of the likelihood and the prior[4,68,69]. On one hand, the word *prior* suggests it must be chosen first; however, without knowledge of the likelihood it makes little sense to start selecting priors - we may not even know the parameters of the model. Thinking of the likelihood and prior as a pair reduces this conceptual tension. It also leads to an explicit way to check our modelling assumptions via generative and predictive modelling[69]. A generative model generates data consistent with the data. The prior has good predictive properties if the *posterior predictive distribution* can predict new data generated from similar experiments. To be explicit, given a likelihood and prior, we can simulate data y. First, sample the parameters of the likelihood, $L(\theta|D)$, from the prior, $p(\theta|\alpha)$, and then given these parameters sample data from the model:

$$\begin{aligned} \tilde{\theta} &\sim p(\theta|\alpha) \\ \tilde{y} &\sim p(y|\tilde{\theta}). \end{aligned} \tag{5}$$

This leads us to define the *prior predictive distribution*:

$$p(\tilde{y}|\alpha) = \int_{\theta} p(\tilde{y}|\theta)p(\theta|\alpha)\,\mathrm{d}\theta. \tag{6}$$

There are a number of key observations. Firstly, the prior predictive distribution has no knowledge of the data, aside from the modelling assumptions of the domain expert. Secondly, the likelihood and prior are now explicitly coupled and so poor modelling choices in either the likelihood or prior will be apparent via the prior predictive. Thirdly, the failure of uniform or uninformative priors as a default is clear, as they will generate unrealistic data.

In our oops example, we model log protein abundance as a linear model of sample type (whether total or oops) and time $(0, 6, 23\text{h})$. Since we are interested in changes in the proportion of protein bound to RNA, we include an interaction effect between time and sample type. We then use Gaussian priors on the coefficients of the effects and an exponential prior on the standard deviation of the Gaussian noise. Formally, the model can be written as

$$\begin{aligned} \log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) \\ \beta &\sim \mathcal{N}(0, 1) \\ \sigma &\sim \mathrm{Exp}(1). \end{aligned} \tag{7}$$

The priors were chosen arbitrarily and we can use a prior predictive check to see whether this leads to a sensible generative model. Figure 2 show a variety of prior predictive checks using different summaries of our observed and simulated data. We see that the our generative model is too diffuse compare with the observed data and produces large deviations beyond what we would expect from a typical proteomics dataset. Hence, it is necessary to explore more prior choices using prior predictive checks. In the accompanying vignette, we show that our inferences can be better calibrated using an exponential prior with rate 4, which corresponds to 1 in 5000 proteins having a standard deviation in their log abundance above 2.5.

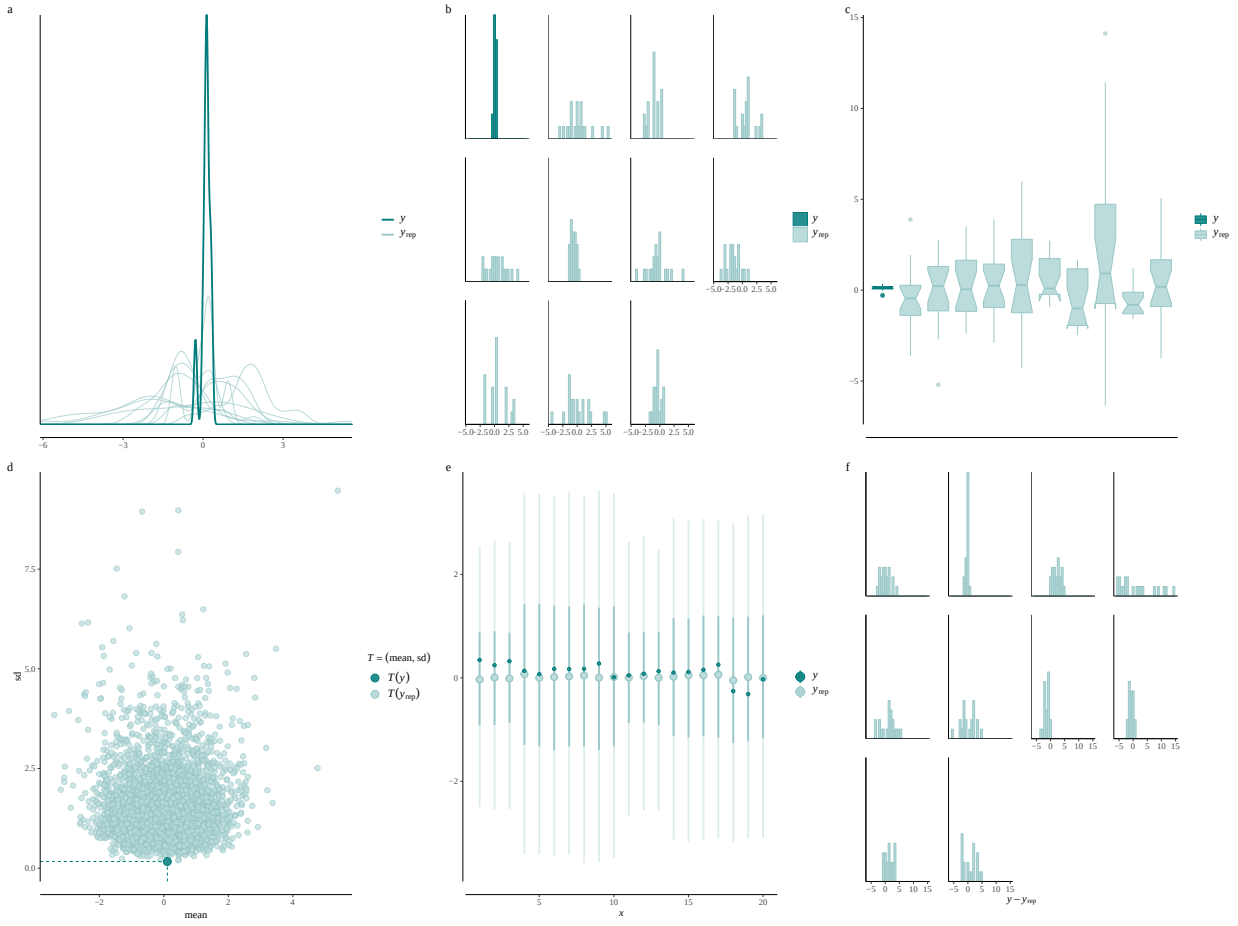

Figure 2: **Prior predictive checks.** Prior predictive checks applied to oops data. y denotes the observed data, whilst y_{rep} denotes the simulated data from the prior predictive distribution. (a) kernel density estimation based checks (b) Histogram based checks (c) boxplot based checks (d) summary statistics checks (e) interval plot based checks (f) error histogram based checks. This figure can be reproduced in the vignette and evaluated for other prior choices.

2.3.3 Predictive modelling

Once our prior and likelihood have seen the data, D, they are updated into the posterior distribution. We can then sample new data by first sampling parameters from the posterior distribution and then again sampling from the likelihood:

$$\begin{aligned} \tilde{\theta} &\sim p(\theta|D, \alpha) \\ \tilde{y} &\sim p(y|\tilde{\theta}). \end{aligned} \tag{8}$$

This leads to the definition of the posterior predictive distribution:

$$p(\tilde{y}|D, \alpha) = \int_\theta p(\tilde{y}|\theta)p(\theta|D, \alpha)\,\mathrm{d}\theta = \int_\theta p(\tilde{y}|\theta)\frac{p(D|\theta, \alpha)p(\theta|\alpha)}{p(D|\alpha)}\mathrm{d}\theta. \tag{9}$$

We have expanded the integrand using Bayes theorem to make a key point explicit: the posterior predictive distribution depends on the likelihood, the prior and the data. This coupling allows us to make a number of observations. A good choice of prior and likelihood leads to good predictive performance and over-fitting can be examined via the posterior predictive distribution.

Having fitted the model to the data, we can perform a posterior predictive check on our inferences. Figure 3 shows a number of posterior predictive checks and that clearly the model has learnt from the data. Visualisation show that samples from the posterior predictive distribution look similar to the observed data. Contrast this with prior predictive checks where the samples from the distribution were very diffuse.

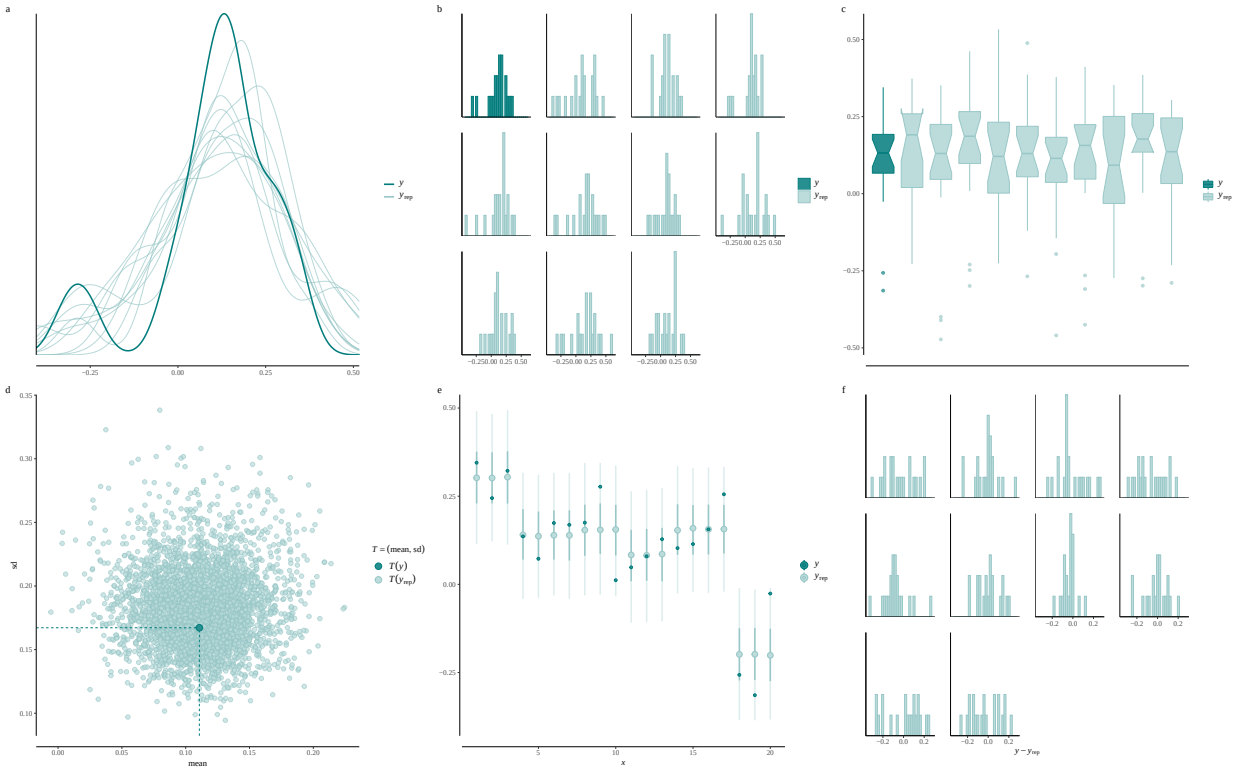

Figure 3: **Posterior predictive checks.** Posterior predictive checks applied to oops data. y denotes the observed data, whilst y_{rep} denotes the simulated data from the prior predictive distribution. (a) kernel density estimation based checks (b) Histogram based checks (c) boxplot based checks (d) summary statistics checks (e) interval plot based checks (f) error histogram based checks. This figure can be reproduced using the vignette and more choices can be explored.

2.3.4 Fitting a model: Bayesian computation

In practice, the integrals and probability distribution required for sufficiently flexible modelling are intractable. We can perform inference in a wide array of models using Markov-chain

Monte Carlo (MCMC) methods[70,71]: including Gibbs sampling[72,73], Metropolis sampling[74], and Hamiltonian Monte Carlo[75]. Bayesian inference can also be performed using sequential Monte Carlo[76] or variational inference[77]. Although the latter can provide a fast approximation of the posterior distribution, it can be arbitrarily inaccurate. Here, we focus on Hamiltonian Monte Carlo, as it forms the basis of modern probabilistic programming languages[78].

Initially, when an MCMC algorithm begins it will "move" towards the posterior distribution producing a "sample" at each iteration. An initial warm-up or burn-in section is require to remove bias due to dependence of the algorithms starting values and to adapt some of the algorithms tuning parameters to provide efficient inference. Once the warm-up section is complete, there is a sampling period which is run until multiple chains have mixed. One measure of mixing chains is $\hat{R}$, which is essentially a measure of between and within chain variance[79]. Current standard practise is that $\hat{R}$ should be close to 1. It is also recommend to visualise trace plots and rank histograms for samples from an MCMC algorithm[80]. Some tools include further diagnostic checks such as divergences but this is beyond the scope of this review[81]. Table 1 highlights some probablistic programming languages that can be used to fit general purpose Bayesian models. For our oops example Bayesian computations are reliable, see accompanying vignette and the supplement.

Packages for Bayesian computation			
Computational tool	language	Inference method	Reference
stan	c++	HMC variant	[78]
brms	R	HMC variant	[82]
MCMCglmm	R	Metropolis/Slice sampling	[83]
PyMC3	Python	HMC variant	[84]
BUGS	BUGS/R	Gibbs sampling	[85]
Edward	Python	Various including variational inference	[86]
Pyro	Python	HMC variant	[87]
Turing.jl	Julia	Various including HMC	[88]

Table 1: **General purpose probabilistic programming languages.** A variety of probabilistic programming languages are available in several languages using modern and efficient inference methods. Amongst these languages, one can fit the vast majority of models used in practice.

2.3.5 Posterior z-scores and contraction

It is often desirable to evaluate the behaviour of a model, and if any model assumptions are preventing us from making sensible inferences. The *posterior z-score* and *posterior contraction* are useful metrics to identify several problems with a model[69]. Let's assume, we have

access to a parameter, θ^*, of the true data generating process. The *posterior z-score* for a parameter is defined as:

$$z_{\text{post}}(\theta|\tilde{y}, \theta^*) = \frac{E_{\text{post}}[\theta|\tilde{y}] - \theta^*}{s_{\text{post}}(\theta|\tilde{y})}, \tag{10}$$

where E_{post} denotes the expectation under the posterior and s_{post} the standard deviation under the posterior. The *posterior contraction* is defined as

$$c(\theta|\tilde{y}) = 1 - \frac{V_{\text{post}}(\theta|\tilde{y})}{V_{\text{prior}}(\theta|\tilde{y})}, \tag{11}$$

where $V_{\text{post/prior}}$ denotes the variance under the posterior/prior. Together these quantities tell us about how the posterior is learning from the data. If the posterior z-score is large and the posterior contraction is small, then the prior modelling conflicts with the true process - we are unable to learn the true parameter well. If the posterior z-score is large and the posterior contraction is close to 1, this suggest we are concentrating on an incorrect part of the probability space and so the model is over-fitting. If the posterior z-score is small and the posterior contraction is also small then the model is poorly informed by the data. The ideal scenario is that posterior contractions are close to 1, and that posterior z-scores are close to 0. This tells us that the data is highly informative and the prior was not biased away from the data generating mechanism. Examples of posterior contractions are shown in the accompanying vignette.

2.3.6 Model selection and averaging

Using probability allows us to select between competing models that may generate the data. Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we can ask for the $P(D|\mathcal{M}_i)$ for $i = 1, 2$. The relative plausibility of two models is referred to as the *Bayes factor* [89],

$$\text{BF}_{12} = \frac{P(D|\mathcal{M}_1)}{P(D|\mathcal{M}_2)} = \frac{P(\mathcal{M}_1|D)}{P(\mathcal{M}_2|D)} \frac{P(\mathcal{M}_2)}{P(\mathcal{M}_1)}. \tag{12}$$

The Bayes factor allows an interpretable and quantitative way to evaluate the relative plausibility of two models, by examining the ratio of the probabilities of the model generating the data. From a brief calculation, we can see that:

$$P(D|\mathcal{M}_1) = \int p(D|\theta_1, \mathcal{M}_1) p(\theta_1|\mathcal{M}_1) \, \mathrm{d}\theta_1, \tag{13}$$

where θ_1 are parameters that parametrise model $\mathcal{M}_1$. Here, we see the dependence of the Bayes Factor on the prior and the implicit assumption that we are evaluating models on their prior predictive performance becomes explicit. Thus, using improper/uninformative priors with Bayes factor would be inappropriate. However, there a further complexities, the most concerning perhaps is that one can inflate the Bayes factor by simply choosing a prior that places probability on unrealistic parts of the parameter space. Typically a uniform prior

would have such an effect. Thus if you are unsure of the veracity of your prior choices, model evaluation may be better using functions of the posterior predictive distributions[69].

We have already seen that one of the key mechanics of Bayesian statistics is the ability to average over quantities, rather than simply taking the best parameters forward. This can also be performed with models using so-called Bayesian model averaging[90]. Let ϕ be a quantity of interest and given models, $\mathcal{M}_i$ $i = 1, ..., n$, we may average them:

$$p(\phi|D) = \sum_{i=1}^{n} p(\phi|D, \mathcal{M}_i)p(\mathcal{M}_i|D). \tag{14}$$

This is the average of the posterior predictive distribution for ϕ under the models considered, weighted by their posterior model probability. If we are interested in the Bayesian model average estimate of a particular parameter, we can compute

$$\hat{\theta} = E_{\text{BMA}}[\theta] = \sum_{i=1}^{n} E_{\mathcal{M}_i}[\theta_i]p(\mathcal{M}_i|D). \tag{15}$$

Given the sensitivity of the Bayes factor to the prior, it is sometimes useful to consider model selection based on the posterior predictive distribution. One example is the log pointwise predictive density (lpd)[91]:

$$\text{lpd} = \sum_{i=1}^{n} \log p(y_i|D) = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|D)\,\mathrm{d}\theta. \tag{16}$$

Furthermore, it is frequently useful to consider an out-of-sample predictive fit via leave-one-out (LOO) cross-validation[91]:

$$\text{lpd}_{\text{LOO}} = \sum_{i=1}^{n} \log p(y_i|D_{-i}), \tag{17}$$

where D_{-i} is data without data point i. This quantity can be efficiently approximated using the LOO package[91]. We note that the above definitions can be adapted to any utility or loss function so that the metric of interest can be characterised.

Returning to our oops example, in our second vignette we develop more complex models of the data. These models include group-level (random) effects for the replicate number and TMT tag used (see model strategies section). The three competing models are
Model 1:

$$\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1).
\end{aligned} \tag{18}$$

Model 2:

$$\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + u_{replicate} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1) \\
u_{replicate} &\sim \mathcal{N}(0, \sigma^2_{replicate}) \\
\log \sigma_{replicate} &\sim \mathcal{T}(3, 0, 0.1)
\end{aligned} \tag{19}$$

Model 3:

$$\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + u_{replicate} + u_{TMT} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1) \\
u_{replicate} &\sim \mathcal{N}(0, \sigma^2_{replicate}) \\
u_{TMT} &\sim \mathcal{N}(0, \sigma^2_{TMT}) \\
\log \sigma_{replicate} &\sim \mathcal{T}(3, 0, 0.1) \\
\log \sigma_{TMT} &\sim \mathcal{T}(3, 0, 0.1)
\end{aligned} \tag{20}$$

Each of the models progresses with more complexity. We compute the posterior model probabilities for each of these examples and find that $P(\mathcal{M}_1|D) = 0.35$, $P(\mathcal{M}_2|D) = 0.48$ and $P(\mathcal{M}_3|D) = 0.17$ (see vignette for more details). This suggest that a group-level effect for replicate is warranted but there is less support for the more complex model 3. Note that because computing the posterior model probabilities includes integration against the prior, these probabilities are automatically penalised for model complexity. See accompanying vignette for further exploration.

2.3.7 Using uncertainty from a Bayesian analysis

Bayesian's quantify uncertainty using probability distributions. Perhaps the most commonly used representation of uncertainty is the credible interval[3]. A credible interval is an interval (a, b) such that a parameter lies within this interval with some probability. For example, we could ask for an interval such that the probability that a protein's log abundance falls between a and b with probability 0.95. In notation used earlier $P(a < \log x < b) = 0.95$. We can see that the interval (a, b) is not unique.

The analogous quantity in frequentist statistics is the confidence interval; however, it is an entirely different concept. This is seen most clearly by asking which part of the constructions are random. For credible intervals, it is the quantity of interest θ that is random and the

interval that is a fixed quantity. Whilst, for a confidence interval the parameter is fixed and the interval is random, since it depends on the randomly observe sample.

However, Bayesian's can report any quantity that can be derived from the posterior distribution or posterior predictive distribution, which in practice can very complex representation of uncertainty. Since summarisation can distort the representation of uncertainty, we recommend reporting the full posterior distribution whenever that is practical.

For our oops example, we are interested in the interaction effects, since these allow us to determine whether the proportion of protein bound to RNA is changing between conditions. In figure 4, we plot the joint distribution of the two interaction effects. We can then ask a number of question of this joint distribution. Some examples include the probability of being positive or negative, the probability of having the opposite signs, the probability that the absolute effects are exceed 0.1, and many more (see accompanying vignette).

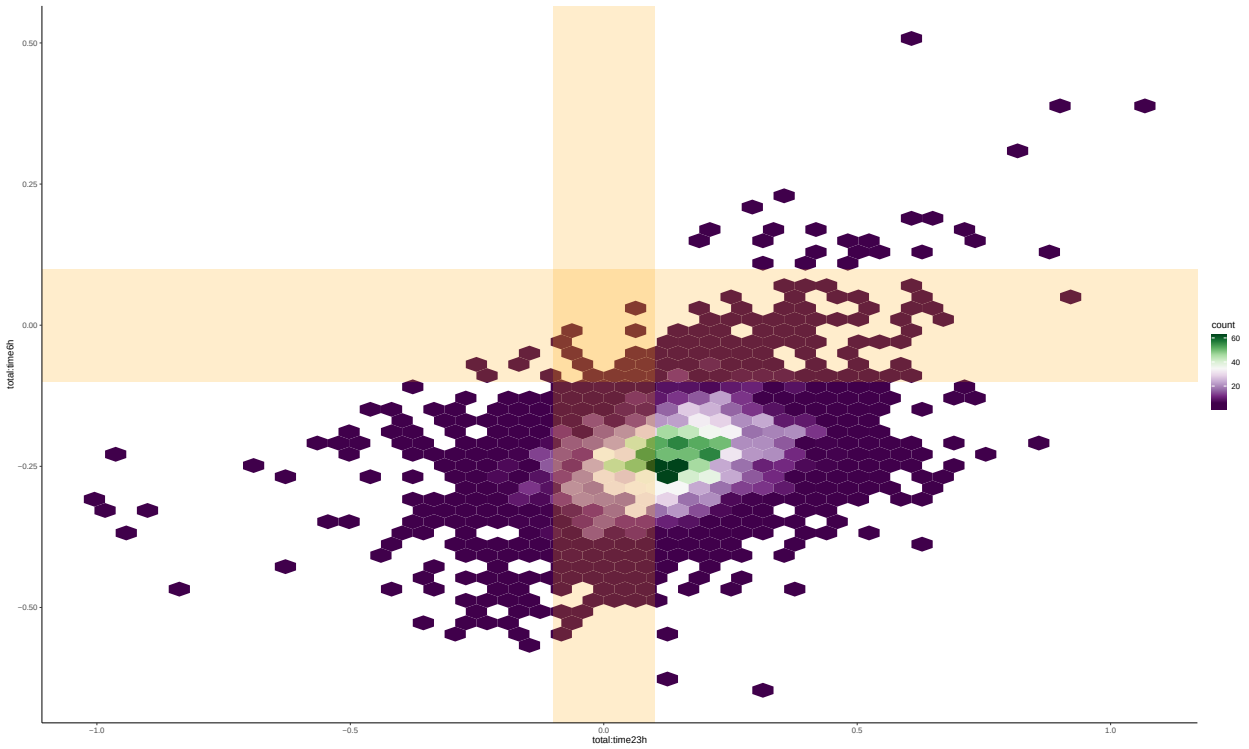

Figure 4: **Joint posterior distribution of interaction effects** Joint posterior distribution of the interaction effect of type with time at six hours and twenty-three hours. The distribution is shown as a 2d histogram with hexagon based density estimation. Orange regions highlight absolute effect sizes less than 0.1. We observe that the density is concentrated outside this region, though is more overlapping at twenty-three hours.

2.4 Modelling strategies

2.4.1 Parametric models

Here, we outline some commonly used modelling strategies and relate them to the proteomics literature. This is not meant to be exhaustive, nor could it be, since there are infinitely many possible models one could specify. One of the most commonly used models is the linear model, where we wish to link a set of predictor to outcomes:

$$y = \beta X + \epsilon. \tag{21}$$

If we choose ϵ to be Gaussian noise, we can write down the model as follows:

$$y \sim \mathcal{N}(\beta X, \sigma^2). \tag{22}$$

There is nothing Bayesian about this model until we specify priors. Remember, the choice of prior should be motivated by generative and predictive modelling and of course the priors should respect the domain of the parameters. Typically, one may start with a Gaussian or Student-t prior on β. The prior on σ could be specified from a variety of probability distributions that respect positivity. Usually recommendations include half-normal, exponential, half-student-t and half-Cauchy depending how confident we are about the scale of the noise [68]. Since protein abundances are positive quantities, it is typical to model them as a log normal distribution

$$\log y \sim \mathcal{N}(\beta X, \sigma^2). \tag{23}$$

If our observed data were counts then it maybe sensible to use Poisson or Negative binomial regression [92]:

$$\begin{aligned} y &\sim \text{Pois}(\lambda) \\ \log(\lambda) &= \beta X \\ \beta &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{24}$$

and

$$\begin{aligned} y &\sim \text{NB}(r, p) \\ \text{logit}(p) &= \beta X \\ \beta &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{25}$$

In each of the above cases, we would have to choose appropriate priors on the model parameters. Again, using the evaluation strategies previously discussed. Many data have an excess of zero's which are not captured by the usual statistical models. Many distribution can be extended to hurdle or zero-inflated models to account for these observations. The distribution of the noise process can be, again, as exotic as needed for the task at hand. Consistent outliers might call for student-t distribution or perhaps the noise itself depends

on some covariates, such as time or spatial location. See Goeminne et al. [93] for an example, albeit non-Bayesian, of a hurdle model applied to proteomics.

Another useful model strategy is to allow parameters at the population-level and group-level, note that these are sometimes referred to as fixed and random effects. For example, a paired t-test is a linear model with grouping specified by the subject or replicate. More complex groupings are allowed, including interactions between groupings and groups that are nested within each other. If β and u are population-level and group-level coefficients with design matrices X and Z, then a log linear (mixed) model would be[94]:

$$\log y = \beta X + uZ + \epsilon. \tag{26}$$

As before the flexibility of the Bayesian analysis, allows you to build any sensible probability distribution on top of this initial model. See Morris et al. [23, 47] for application of mixed-models to proteomics data, as well as the accompanying vignette.

Another useful modelling strategy is mixture models, which occurs frequently in the context of clustering and classification[95]. The mixture model assume that data arises from different components each with the same parametric density with different parameters:

$$\begin{aligned} y_i|z_i, \theta &\sim F(\theta_{z_i}) \\ z_i|\pi &\sim \text{cat}(\pi) \\ \pi|\alpha &\sim \text{Dir}(\alpha) \\ \theta &\sim p(\theta). \end{aligned} \tag{27}$$

The priors and the likelihood can be chosen based on the specific application at hand and the workflow recommendations can be applied. It is often insightful to write, using the law of total probability, the mixture model as

$$p(y_i) = \sum_{k=1}^{K} \pi_k p(y_i|\theta_k). \tag{28}$$

Note that because a Dirichlet prior is placed on π, the entries must all be non-negative and sum to unity. Hence, the entries of π can be interpreted as weights. The data cluster by being associated to the component density which fits those observations through the variables z_i. Examples of mixture models applied to proteomics include Chung et al. [40], Crook et al. [61, 62]

2.4.2 Non-parametric models

In contrast to parametric models, non-parametric models allow more parameters as more data is observed. Phrased another way, in a parametric model there are finitely many parameters, whilst in a non-parametric model there are infinitely many such parameters. This makes non-parametric models more flexible; however, to avoid the over-fitting concerns raised in earlier sections, we ought to be prudent with our choice of priors. One of the most

popular non-parametric model is the Gaussian process (GP), which can be used to model functions f. Suppose we observe data $\{(x_i, y_i)_{i=1,\ldots,n}\}$, we wish to find a function f such that $f(x_i)$ models y_i. Let us assume a Gaussian regression set-up, using a *Gaussian process prior* to model f:

$$\begin{aligned} y &\sim \mathcal{N}(f, \sigma^2) \\ f &\sim \mathcal{GP}(m, C). \end{aligned} \tag{29}$$

The Gaussian process is a distribution over *functions* that is uniquely characterised by its mean and covariance functions. The choice of mean and covariance functions are modelling choices to be made by the domain expert. Typically, the covariance function is parametrised by some parameters $C = C(\theta)$ and we can also place priors on these parameters so that $\theta \sim p(\theta)$. Again, these modelling choices can be evaluated using prior/posterior predictive checks. We refer to several discussion on choose priors for Gaussian processes[96,97,98,99,100]. For applications of Gaussian process to proteomics data see Maboudi Afkham et al. [55], Crook et al. [62], Fang et al. [66], Shin et al. [101].

The other non-parametric model that is frequently used is the Dirichlet process[102,103]. Dirichlet processes are a popular tool for modelling data with parameter repetitions. For example, when we cluster data, all observation associated with cluster 1 share the same parameter θ_1. The Dirichlet process is defined using a base distribution G and a concentration parameter α and is written $\mathrm{DP}(G, \alpha)$. For example, suppose that $G = \mathcal{N}(0, 1)$, then we can simulate from the Dirichlet process as follows. For any $i \geq 1$, with probability $\frac{\alpha}{\alpha+i-1}$ sample $x_i \sim \mathcal{N}(0, 1)$ and with probability $\frac{n_x}{\alpha+i-1}$ let $x_i = x$, where n_x is the number of previous observations of x. This means, if we have already observed a value, then we are increasingly likely to observe it in the future. This property is sometime referred to as the "rich get richer property".

The Dirichlet process allows us to work with mixture models with infinitely many components, which is useful for characterising the uncertainty in the number of components. Once we have a sensible parametric likelihood for the observations $F(\theta_i)$, the Dirichlet process can be used as a prior to construct the Dirichlet process mixture model:

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|P &\sim P \\ P|\alpha, G &\sim \mathrm{DP}(G, \alpha). \end{aligned} \tag{30}$$

Since P will be discrete, the set$\{\theta_i\}_{i=1,\ldots,n}$ will contain repetitions. This allows us to think of this model as a mixture model, where the groups of parameters define the components. Extensions are available[104,105] and MCMC algorithms for fitting these models can be found in Neal [106]. For applications of Dirichlet processes to proteomics data see Claassen et al. [39] and Choi et al. [65].

3 Discussion

Despite Bayesian statistics offering a powerful and flexible framework for performing proteomics data analysis, only a few problems have yet been tackled using this methodology. Even when Bayesian statistics has been applied, the methodology has not made complete use of the information available from such an analysis. Many analysis have simply resorted to proxies from frequentist based approaches. One of the key advantages of the Bayesian approach is to be able to jointly model several quantities and provide uncertainty estimates in any parameters. Another advantage of Bayesian statistics is that it makes modelling assumption explicit; hence, it becomes clear how the models can be improved and what is the extent of their limitations.

Here, we have summarised key modelling ideas in Bayesian statistics starting with the workflow. We have highlighted that the Bayesian workflow has a consistent approach to model building, model criticism and evaluation grounded in probability theory. Using a case study, we have provided a workflow for developing a Bayesian model for Organic Orthogonal Phase Separation (oops) data. We then proceeded to describe and illustrate common modelling strategies to help proteomics researchers understand key models in the literature and link them to current methods used in the literature.

Mass spectrometry-based proteomics appears to have resisted uptake on Bayesian methods for various reasons. These include, but are not limited to, lack of familiarity with the workflow and tools available, lack of compelling examples in literature, and lack of desire to invest in bespoke model development. We hope that this review goes some way in removing some of these barriers to applying and understand Bayesian methods.

References

[1] Wilkie, D. *Nature* **1974**, *251*, 601–602.

[2] Tversky, A.; Kahneman, D. *science* **1974**, *185*, 1124–1131.

[3] Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian data analysis*; Chapman and Hall/CRC, 1995.

[4] Gelman, A.; Vehtari, A.; Simpson, D.; Margossian, C. C.; Carpenter, B.; Yao, Y.; Kennedy, L.; Gabry, J.; Bürkner, P.-C.; Modrák, M. *arXiv preprint arXiv:2011.01808* **2020**,

[5] Schad, D. J.; Betancourt, M.; Vasishth, S. *Psychological methods* **2021**, *26*, 103.

[6] Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. *Analytical and bioanalytical chemistry* **2007**, *389*, 1017–1031.

[7] Mathieson, T.; Franken, H.; Kosinski, J.; Kurzawa, N.; Zinn, N.; Sweetman, G.; Poeckel, D.; Ratnu, V. S.; Schramm, M.; Becher, I., et al. *Nature communications* **2018**, *9*, 1–10.

[8] Huttlin, E. L.; Ting, L.; Bruckner, R. J.; Gebreab, F.; Gygi, M. P.; Szpyt, J.; Tam, S.; Zarraga, G.; Colby, G.; Baltier, K., et al. *Cell* **2015**, *162*, 425–440.

[9] Mateus, A.; Kurzawa, N.; Becher, I.; Sridharan, S.; Helm, D.; Stein, F.; Typas, A.; Savitski, M. M. *Molecular systems biology* **2020**, *16*, e9232.

[10] Schopper, S.; Kahraman, A.; Leuenberger, P.; Feng, Y.; Piazza, I.; Müller, O.; Boersema, P. J.; Picotti, P. *Nature protocols* **2017**, *12*, 2391–2410.

[11] Masson, G. R.; Burke, J. E.; Ahn, N. G.; Anand, G. S.; Borchers, C.; Brier, S.; Bou-Assaf, G. M.; Engen, J. R.; Englander, S. W.; Faber, J., et al. *Nature methods* **2019**, *16*, 595–602.

[12] Gessel, M. M.; Norris, J. L.; Caprioli, R. M. *Journal of proteomics* **2014**, *107*, 71–82.

[13] Geladaki, A.; Britovšek, N. K.; Breckels, L. M.; Smith, T. S.; Vennard, O. L.; Mulvey, C. M.; Crook, O. M.; Gatto, L.; Lilley, K. S. *Nature communications* **2019**, *10*, 1–15.

[14] Barylyuk, K.; Koreny, L.; Ke, H.; Butterworth, S.; Crook, O. M.; Lassadi, I.; Gupta, V.; Tromer, E.; Mourier, T.; Stevens, T. J., et al. *Cell host & microbe* **2020**, *28*, 752–766.

[15] Toby, T. K.; Fornelli, L.; Kelleher, N. L. *Annual review of analytical chemistry* **2016**, *9*, 499–519.

[16] Nightingale, K.; Lin, K.-M.; Ravenhill, B. J.; Davies, C.; Nobre, L.; Fielding, C. A.; Ruckova, E.; Fletcher-Etherington, A.; Soday, L.; Nichols, H., et al. *Cell host & microbe* **2018**, *24*, 447–460.

[17] Johnson, D. T.; Di Stefano, L. H.; Jones, L. M. *Journal of Biological Chemistry* **2019**, *294*, 11969–11979.

[18] Queiroz, R. M.; Smith, T.; Villanueva, E.; Marti-Solano, M.; Monti, M.; Pizzinga, M.; Mirea, D.-M.; Ramakrishna, M.; Harvey, R. F.; Dezi, V., et al. *Nature biotechnology* **2019**, *37*, 169–178.

[19] Talts, S.; Betancourt, M.; Simpson, D.; Vehtari, A.; Gelman, A. *arXiv preprint arXiv:1804.06788* **2018**,

[20] Crook, O. M.; Breckels, L. M.; Lilley, K. S.; Kirk, P. D.; Gatto, L. *F1000Research* **2019**, *8*.

[21] Ni, Y.; Stingo, F. C.; Ha, M. J.; Akbani, R.; Baladandayuthapani, V. *Journal of the American Statistical Association* **2019**, *114*, 48–60.

[22] Maity, A. K.; Bhattacharya, A.; Mallick, B. K.; Baladandayuthapani, V. *Biometrics* **2020**, *76*, 316–325.

[23] Morris, J. S.; Baladandayuthapani, V.; Herrick, R. C.; Sanna, P.; Gutstein, H. *The annals of applied statistics* **2011**, *5*, 894.

[24] Phillips, A.; Unwin, R. D.; Hubbard, S.; Dowsey, A. *Statistical methods for proteomics* **2021**,

[25] The, M.; Käll, L. *Journal of proteome research* **2021**, *20*, 2062–2068.

[26] The, M.; Käll, L. *Molecular & cellular Proteomics* **2019**, *18*, 561–570.

[27] Santra, T.; Delatola, E. I. *Scientific reports* **2016**, *6*, 1–10.

[28] Peshkin, L.; Gupta, M.; Ryazanova, L.; Wühr, M. *Molecular & Cellular Proteomics* **2019**, *18*, 2108–2120.

[29] Millikin, R. J.; Shortreed, M. R.; Scalf, M.; Smith, L. M. *Journal of proteome research* **2020**, *19*, 1975–1981.

[30] Serang, O.; Cansizoglu, A. E.; Käll, L.; Steen, H.; Steen, J. A. *Journal of proteome research* **2013**, *12*, 4556–4565.

[31] O'Brien, J. J.; O'Connell, J. D.; Paulo, J. A.; Thakurta, S.; Rose, C. M.; Weekes, M. P.; Huttlin, E. L.; Gygi, S. P. *Journal of proteome research* **2018**, *17*, 590–599.

[32] Serang, O.; Moruz, L.; Hoopmann, M. R.; Käll, L. *Journal of proteome research* **2012**, *11*, 5586–5591.

[33] Carvalho, P. C.; Fischer, J. S.; Perales, J.; Yates, J. R.; Barbosa, V. C.; Bareinboim, E. *Bioinformatics* **2011**, *27*, 275–276.

[34] Jow, H.; Boys, R. J.; Wilkinson, D. J. *Statistical applications in genetics and molecular biology* **2014**, *13*, 531–551.

[35] O'Brien, J. J.; Gunawardena, H. P.; Paulo, J. A.; Chen, X.; Ibrahim, J. G.; Gygi, S. P.; Qaqish, B. F. *The annals of applied statistics* **2018**, *12*, 2075.

[36] Chen, S. S.; Deutsch, E. W.; Yi, E. C.; Li, X.-j.; Goodlett, D. R.; Aebersold, R. *Journal of proteome research* **2005**, *4*, 2174–2184.

[37] Halloran, J. T.; Bilmes, J. A.; Noble, W. S. *Journal of proteome research* **2016**, *15*, 2749–2759.

[38] Lewis, N. H.; Hitchcock, D. B.; Dryden, I. L.; Rose, J. R. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **2018**, *67*, 1207–1236.

[39] Claassen, M.; Aebersold, R.; Buhmann, J. M. *Bioinformatics* **2009**, *25*, i154–i160.

[40] Chung, C.; Emili, A.; Frey, B. J. *Bioinformatics* **2013**, *29*, 821–829.

[41] Webb-Robertson, B.-J. M.; Matzke, M. M.; Datta, S.; Payne, S. H.; Kang, J.; Bramer, L. M.; Nicora, C. D.; Shukla, A. K.; Metz, T. O.; Rodland, K. D., et al. *Molecular & cellular proteomics* **2014**, *13*, 3639–3646.

[42] Lim, M. Y.; O'Brien, J.; Paulo, J. A.; Gygi, S. P. *Journal of proteome research* **2017**, *16*, 4217–4226.

[43] Shteynberg, D. D.; Deutsch, E. W.; Campbell, D. S.; Hoopmann, M. R.; Kusebauch, U.; Lee, D.; Mendoza, L.; Midha, M. K.; Sun, Z.; Whetton, A. D., et al. *Journal of proteome research* **2019**, *18*, 4262–4272.

[44] Mallikarjun, V.; Richardson, S. M.; Swift, J. *Journal of proteome research* **2020**, *19*, 2167–2184.

[45] Hernández, B.; Pennington, S. R.; Parnell, A. C. *EuPA Open Proteomics* **2015**, *9*, 54–64.

[46] Morris, J. S.; Brown, P. J.; Baggerly, K. A.; Coombes, K. R. **2006**,

[47] Morris, J. S.; Brown, P. J.; Herrick, R. C.; Baggerly, K. A.; Coombes, K. R. *Biometrics* **2008**, *64*, 479–489.

[48] Liao, H.; Moschidis, E.; Riba-Garcia, I.; Zhang, Y.; Unwin, R. D.; Morris, J. S.; Graham, J.; Dowsey, A. W. **2014**, 1332–1335.

[49] Hwang, D.; Zhang, N.; Lee, H.; Yi, E.; Zhang, H.; Lee, I. Y.; Hood, L.; Aebersold, R. *Bioinformatics* **2008**, *24*, 2641–2642.

[50] Harris, K.; Girolami, M.; Mischak, H. **2009**, 137–149.

[51] Kuschner, K. W.; Malyarenko, D. I.; Cooke, W. E.; Cazares, L. H.; Semmes, O. J.; Tracy, E. R. *BMC bioinformatics* **2010**, *11*, 1–10.

[52] Deng, X.; Geng, H.; Ali, H. H. *Cancer informatics* **2007**, *3*, 117693510700300001.

[53] Liu, Y.; Li, M.; Morris, J. S. *The Annals of Applied Statistics* **2020**, *14*, 521–541.

[54] Chen, A. T.; Franks, A.; Slavov, N. *PLoS computational biology* **2019**, *15*, e1007082.

[55] Maboudi Afkham, H.; Qiu, X.; The, M.; Käll, L. *Bioinformatics* **2017**, *33*, 508–513.

[56] Saltzberg, D. J.; Broughton, H. B.; Pellarin, R.; Chalmers, M. J.; Espada, A.; Dodge, J. A.; Pascal, B. D.; Griffin, P. R.; Humblet, C.; Sali, A. *The Journal of Physical Chemistry B* **2017**, *121*, 3493–3501.

[57] LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L. *Journal of proteome research* **2014**, *13*, 3231–3240.

[58] Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K.; Benesch, J. L.; Robinson, C. V. *Analytical chemistry* **2015**, *87*, 4370–4376.

[59] Kostelic, M.; Zak, C.; Liu, Y.; Chen, V.; Wu, Z.; Sivinski, J.; Chapman, E.; Marty, M. **2021**,

[60] Christopher, J. A.; Stadler, C.; Martin, C. E.; Morgenstern, M.; Pan, Y.; Betsinger, C. N.; Rattray, D. G.; Mahdessian, D.; Gingras, A.-C.; Warscheid, B., et al. *Nature Reviews Methods Primers* **2021**, *1*, 1–24.

[61] Crook, O. M.; Mulvey, C. M.; Kirk, P. D.; Lilley, K. S.; Gatto, L. *PLoS computational biology* **2018**, *14*, e1006516.

[62] Crook, O. M.; Lilley, K. S.; Gatto, L.; Kirk, P. D. *arXiv preprint arXiv:1903.02909* **2019**,

[63] Crook, O. M.; Geladaki, A.; Nightingale, D. J.; Vennard, O. L.; Lilley, K. S.; Gatto, L.; Kirk, P. D. *PLoS computational biology* **2020**, *16*, e1008288.

[64] Crook, O. M.; Davies, C. T.; Gatto, L.; Kirk, P. D.; Lilley, K. S. *bioRxiv* **2021**,

[65] Choi, H.; Kim, S.; Gingras, A.-C.; Nesvizhskii, A. I. *Molecular systems biology* **2010**, *6*, 385.

[66] Fang, S.; Kirk, P. D.; Bantscheff, M.; Lilley, K. S.; Crook, O. M. *Communications biology* **2021**, *4*, 1–15.

[67] Schmiesing, J. A.; Gregson, H. C.; Zhou, S.; Yokomori, K. *Molecular and cellular biology* **2000**, *20*, 6996–7006.

[68] Gelman, A.; Simpson, D.; Betancourt, M. *Entropy* **2017**, *19*, 555.

[69] Betancourt, M. *https://github.com/betanalpha/knitr_case_studies/tree/master/principled_bayesian_workflow* **2021**,

[70] Gilks, W. R.; Richardson, S.; Spiegelhalter, D. *Markov chain Monte Carlo in practice*; CRC press, 1995.

[71] Brooks, S.; Gelman, A.; Jones, G.; Meng, X.-L. **2011**,

[72] Smith, A. F.; Roberts, G. O. *Journal of the Royal Statistical Society: Series B (Methodological)* **1993**, *55*, 3–23.

[73] Gelfand, A. E. *Journal of the American statistical Association* **2000**, *95*, 1300–1304.

[74] Robert, C. P.; Casella, G. *Monte Carlo Statistical Methods*; Springer, 1999; pp 231–283.

[75] Hoffman, M. D.; Gelman, A., et al. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.

[76] Del Moral, P.; Doucet, A.; Jasra, A. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2006**, *68*, 411–436.

[77] Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. *Journal of the American statistical Association* **2017**, *112*, 859–877.

[78] Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. *Journal of statistical software* **2017**, *76*, 1–32.

[79] Vehtari, A.; Gelman, A.; Simpson, D.; Carpenter, B.; Bürkner, P.-C. *arXiv preprint arXiv:1903.08008* **2019**,

[80] Gabry, J.; Simpson, D.; Vehtari, A.; Betancourt, M.; Gelman, A. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **2019**, *182*, 389–402.

[81] Betancourt, M. *arXiv preprint arXiv:1701.02434* **2017**,

[82] Bürkner, P.-C. *Journal of statistical software* **2017**, *80*, 1–28.

[83] Hadfield, J. D. *Journal of statistical software* **2010**, *33*, 1–22.

[84] Salvatier, J.; Wiecki, T. V.; Fonnesbeck, C. *PeerJ Computer Science* **2016**, *2*, e55.

[85] Lunn, D.; Spiegelhalter, D.; Thomas, A.; Best, N. *Statistics in medicine* **2009**, *28*, 3049–3067.

[86] Tran, D.; Kucukelbir, A.; Dieng, A. B.; Rudolph, M.; Liang, D.; Blei, D. M. *arXiv preprint arXiv:1610.09787* **2016**,

[87] Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N. D. *The Journal of Machine Learning Research* **2019**, *20*, 973–978.

[88] Ge, H.; Xu, K.; Ghahramani, Z. Turing: A language for flexible probabilistic inference. International conference on artificial intelligence and statistics. 2018; pp 1682–1690.

[89] Kass, R. E.; Raftery, A. E. *Journal of the american statistical association* **1995**, *90*, 773–795.

[90] Raftery, A. E.; Madigan, D.; Hoeting, J. A. *Journal of the American Statistical Association* **1997**, *92*, 179–191.

[91] Vehtari, A.; Gelman, A.; Gabry, J. *Statistics and computing* **2017**, *27*, 1413–1432.

[92] Lawless, J. F. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **1987**, 209–225.

[93] Goeminne, L. J.; Sticker, A.; Martens, L.; Gevaert, K.; Clement, L. *Analytical chemistry* **2020**, *92*, 6278–6287.

[94] Bates, D.; Mächler, M.; Bolker, B.; Walker, S. *arXiv preprint arXiv:1406.5823* **2014**,

[95] McLachlan, G. J.; Lee, S. X.; Rathnayake, S. I. *Annual review of statistics and its application* **2019**, *6*, 355–378.

[96] Berger, J. O.; De Oliveira, V.; Sansó, B. *Journal of the American Statistical Association* **2001**, *96*, 1361–1374.

[97] Paulo, R., et al. *The Annals of Statistics* **2005**, *33*, 556–582.

[98] De Oliveira, V. *Canadian Journal of Statistics* **2007**, *35*, 283–301.

[99] van der Vaart, A. W.; van Zanten, J. H., et al. *The Annals of Statistics* **2009**, *37*, 2655–2675.

[100] Fuglstad, G.-A.; Simpson, D.; Lindgren, F.; Rue, H. *Journal of the American Statistical Association* **2019**, *114*, 445–452.

[101] Shin, J. J.; Crook, O. M.; Borgeaud, A. C.; Cattin-Ortolá, J.; Peak-Chew, S. Y.; Breckels, L. M.; Gillingham, A. K.; Chadwick, J.; Lilley, K. S.; Munro, S. *Nature communications* **2020**, *11*, 1–13.

[102] Ferguson, T. S. *The annals of statistics* **1973**, 209–230.

[103] Antoniak, C. E. *The annals of statistics* **1974**, 1152–1174.

[104] Teh, Y. W.; Jordan, M. I.; Beal, M. J.; Blei, D. M. *Journal of the american statistical association* **2006**, *101*, 1566–1581.

[105] Rodriguez, A.; Dunson, D. B.; Gelfand, A. E. *Journal of the American statistical Association* **2008**, *103*, 1131–1154.

[106] Neal, R. M. *Journal of computational and graphical statistics* **2000**, *9*, 249–265.