Does Interactive Conditioning Help Users Better Understand the Structure of Probabilistic Models?

Evdoxia Taka, Sebastian Stein, and John H. Williamson

Abstract—Despite growing interest in probabilistic modeling approaches and availability of learning tools, people are hesitant to use them. There is a need for tools to communicate probabilistic models more intuitively and help users build, validate, use effectively or trust probabilistic models. We focus on visual representations of probabilistic models and introduce the Interactive Pair Plot (IPP) for visualization of a model's uncertainty, a scatter plot matrix of a probabilistic model allowing interactive conditioning on the model's variables. We investigate whether the use of interactive conditioning in a scatter plot matrix of a model helps users better understand variables' relations. We conducted a user study and the findings suggest that improvements in the understanding of the interaction group are the most pronounced for more exotic structures, such as hierarchical models or unfamiliar parameterizations, in comparison to the understanding of the static group. As the detail of the inferred information increases, interactive conditioning does not lead to considerably longer response times. Finally, interactive conditioning improves participants' confidence about their responses.

Index Terms—Brushing-and-linking, empirical study, interactive conditioning, prior distribution, probabilistic models, scatter plot matrix.

1 Introduction

Probabilistic modeling is a form of statistical modeling that has increased in popularity lately especially in the context of Bayesian analysis. The emergence of Probabilistic Programming Languages (PPLs) (e.g. Stan, PyMC) made probabilistic modeling accessible to a broader audience. Despite the growing interest, these methods are not widely adopted. Non-experienced researchers who conduct experiments and analyze data do not feel confident to use such methods for the analysis [1, 2] even when they have access to learning and exploration tools [3]. Users, who need to rely on such models to do their job, might find it difficult to understand their structure. Decision-makers with moderate statistical background might make uninformed and potentially risky decisions because they cannot understand the effect of intervening on a variable upon other variables in a model. The mathematical definition of probabilistic models can be complex, unintuitive and hard to understand even for more experienced users [4].

Understanding the relations among variables in a probabilistic model given definitions of the model in textual languages or graphs [5–8] (Fig. 1) is very much dependent on users' statistical knowledge. For example, variable `b` in Model 1 (Fig. 1a) controls the mean value of variable `temperature`. Increasing `b`'s value would increase `temperature`'s mean value. In models where relations are governed by more complex statistical or mathematical associations, it requires good statistical knowledge to tell what the effect of a variable on others would be. There is a need for tools to communicate variables' relations in probabilistic models more intuitively and help users build, validate, use effectively or trust probabilistic models.

Variables' relations in a probabilistic model can be visualized through visualizations of variables' inherent uncertainty. Scatter plot matrices present variables' pairwise distributions conveying existing correlations. IPME [9] is a graphical representation with nodes corresponding to models' variables and showing the KDE (Kernel Density Estimation) plot of the variable (Fig. 2l-s). It uses interactive conditioning implemented as a brushing-and-linking interactivity on KDE plots to enable a form of "sensitivity analysis" of the variables and reveal their relations.

Various visualization designs were explored in the literature to facilitate reasoning about unintuitive mathematical concepts like uncertainty [10–12] and Bayes' rule [13–19]. Brushing-and-linking [20–22], although it is often present in scatter plot matrices to help with the exploration of multidimensional data, is rarely evaluated for its efficiency with real test subjects [23, 24].

In this paper, we introduce an *interactive pair plot (IPP)*[1] which is a classical scatter plot matrix that integrates IPME's interactive conditioning with some additional sample highlighting. We conducted a user study to investigate whether users make better inferences about variables' relations presented in a scatter plot matrix when they use interactive conditioning in comparison to simply observing a static scatter plot matrix. We focused on which levels of detail of variables' relations (Section 2.2) and which probabilistic model designs (e.g. hierarchical structures, complex parameterizations of variables' distributions) (Section 6.1) inter-

• *E. Taka, S. Stein, and J. H. Williamson are with the School of Computing Science, University of Glasgow, UK. E-mail: e.taka.1@research.gla.ac.uk, {johnh.williamson, sebastian.stein}@glasgow.ac.uk*

Manuscript received February 22, 2022; revised August 23, 2022.

1. https://github.com/evdoxiataka/ipme/releases/tag/ipp

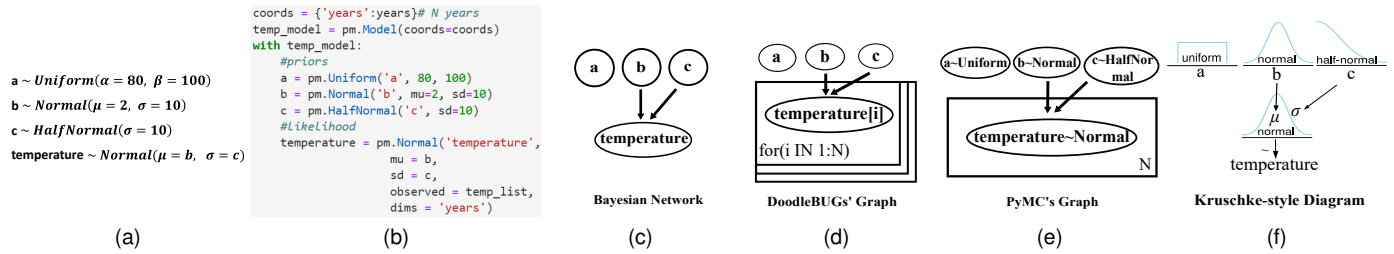


Fig. 1. Visual representations of Model 1 of user study. **Definitions in Textual Languages**: (a) Probabilistic statements. For example, the first statement reads: "Random variable a follows (∼) a uniform probability distribution with the lower bound α = 80 and the upper bound β = 100". (b) PPL code (PyMC) of model for Bayesian inference. A likelihood is defined for the observed variable temperature to account for the list of N observed temperatures for a set of years. **Graphs**: (c)-(f) Transcriptions of model in various graph types.

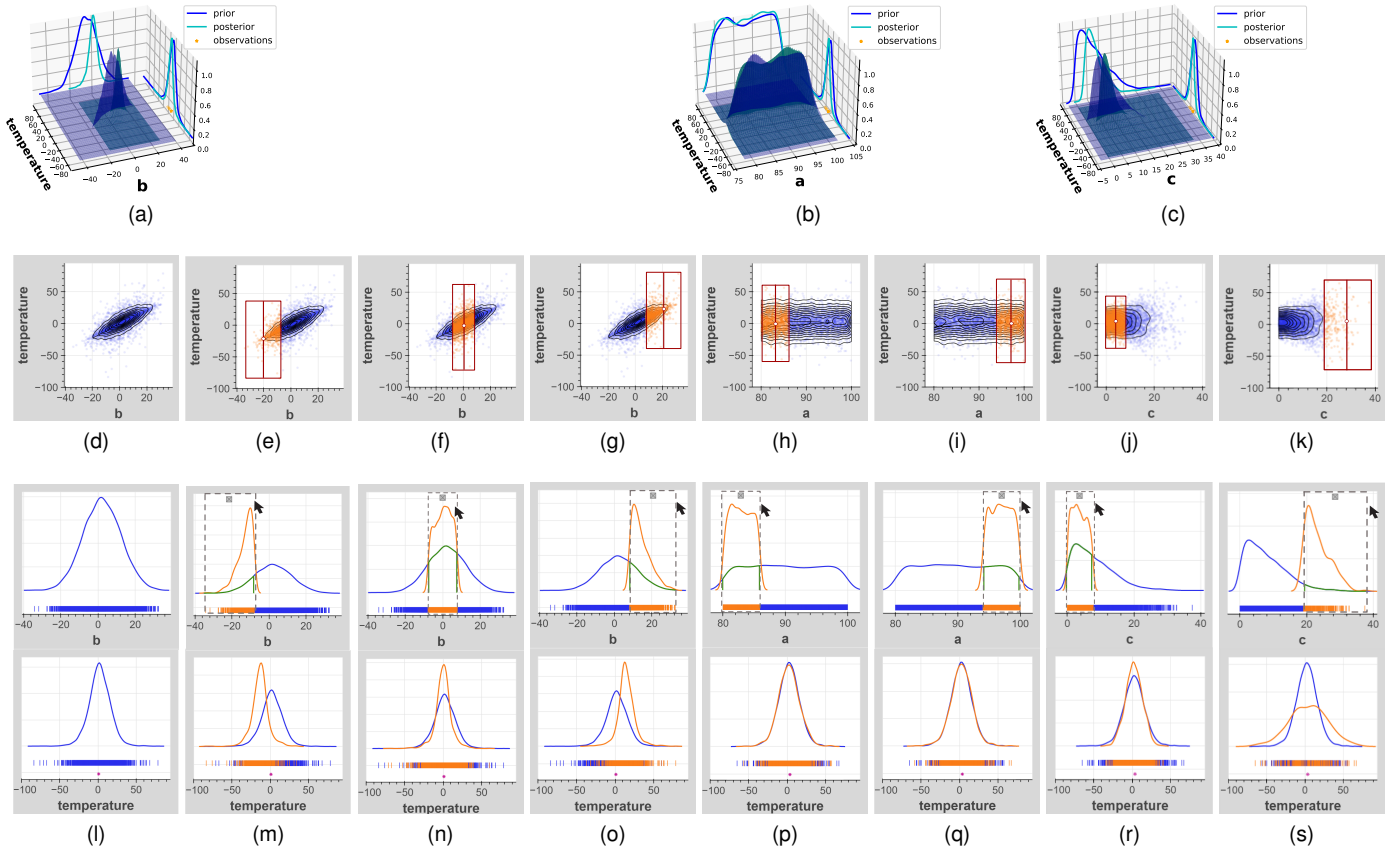


Fig. 2. Visualizations of variables' relations in Model 1 of user study. **Joint & Marginal Distributions**: (a)-(c) The prior and posterior joint (3D surface plots) and marginal distributions (line plots on cube faces) of variables temperature, and b, a, or c, respectively. The yellow stars represent the observations in temp_list. **Scatter Plots**: (d)-(k) Samples and contours of variables' pairwise prior joint distributions. Conditioning facilitates the interpretation of scatter plots' shape. For example, conditioning on b in sequential increased ranges in (e)-(g), increases the mean value (white dot) of temperature's distribution. **Interactive Conditioning with IPME**: (l)-(s) IPME-like representation. Interactive conditioning is applied on the prior marginal distributions of b, a, or c and the conditional marginal distributions are drawn (in orange).

active conditioning can be more beneficial for. We used IPP as the visualization instance and measured participants' accuracy, response times, and confidence.

2 Background

Illustrations in Fig. 1 and 2 accompany the text in Sections 2 and 3. Model 1 of the user study is used as a unifying example in these figures to present various representations of the model with varying levels of information.

2.1 Probabilistic Models

Probabilistic models consist of *random variables* that each follow a probability distribution. Model 1 in Fig. 1a consists of the a, b, c, and temperature random variables. Variable temperature follows a normal distribution having two parameters, the μ and σ. The probability density function (pdf) of a normal distribution is a function of the value the random variable can take given the parameters; $f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. The parameters of a distribution can be random variables, as well. The μ and σ parameters of temperature are set by the b and c random variables.

The probability distributions of the random variables in a probabilistic model are the *marginal distributions* of the model's multi(k)-variate *joint distribution*. In Model 1 $k = 4$.

Probabilistic models' variables are either *observed*, referring to directly observed or measured variables, or *latent*, referring to unobserved hidden variables. In Model 1, `temperature` is an observed variable and `a`, `b`, and `c` are latent variables. This categorization is important in Bayesian analysis. The model's distribution called *prior*, because it encodes the prior knowledge and experience (e.g. possible value ranges) before seeing any observation, gets updated to reflect the posterior beliefs about the model's variables in the light of observed data. The model's distribution after this update is called *posterior distribution*. Fig. 2a presents the prior (in blue color) and posterior (in cyan color) marginal (line plots on cube faces) and joint (3D surface plots) distributions of `temperature` and `b` variables. The prior and posterior distributions of an observed variable are called *predictive*, because samples drawn from them form possible data-sets before or after observing the data.

To estimate the posterior distribution of a model, a likelihood function of the probability distribution of the observed variables is defined to account for the observed data, and Bayes' rule is applied. The likelihood is a function of the distribution's parameters. In the definition of Model 1 in PyMC code in Fig. 1b, a normal likelihood is defined for the `temperature` observed variable. The mathematical transcription of the PPL-defined likelihood for `temperature` would be $\mathcal{L}(\mu, \sigma|\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\frac{\mathrm{x}-\mu}{\sigma})^2}$, where x is set by the list of observations `temp_list`.

2.2 Relations of Probabilistic Models' Variables

Two variables in a probabilistic model are related when one is used for setting the distribution's random-valued (non-fixed) parameters of the other. We propose to define three levels of detail, from lowest to highest, to characterize these relations among variables in a model. We use these levels in our user study to evaluate users' understanding in the different levels of detail as we explain in Section 5.

- L1 **Existence.** Are variables related? Variable `temperature` is related to `b` and `c`, but not to `a` in Model 1 (Fig. 1a).
- L2 **Polarity.** What is the sign (positive or negative) of the polarity of the effect one variable has on another? The μ parameter of `temperature`'s distribution increases when the value of `b` increases in Model 1 (positive polarity), while it would decrease if it was set equal to `-b` (negative polarity).
- L3 **Quantification.** How are variables related? A relation is quantified by the specific statistical associations (which parameters of a distribution are affected by a variable) and formula (mathematical transformation or equation) that sets it. Variable `b` sets `temperature`'s μ parameter through a simple assignment. A transformation $\exp(\mathtt{b})$ or an equation $5 - 2 * \mathtt{b}$ could be more complex ways to do so.

The following section reviews existing visual representations of probabilistic models and visualizations of variables' relations, and explains which of these levels of detail could be retrieved from them, and how.

3 Related Work

3.1 Visual Representations of Probabilistic Models

Textual Language Definitions. Variables' relations (L1) and their quantification (L3) are retrievable through an at-a-glance observation of the model's definition in probabilistic statements (Fig. 1a) or PPL code (Fig. 1b). Retrieving the polarity of variables' effects (L2) is dependent on the ability of the user to interpret the mathematical details.

Graphs. Graphs can hide the mathematical details of probabilistic models, while preserving some structural information. Bayesian networks (Fig. 1c) are informationally minimal having *nodes* corresponding to model's variables, and *edges* (directed arrows) from one variable to another indicating the direction of the relation. Some PPLs produce more informed graphs like the DoodleBUGs' graph [6] where nodes contain information about variables' dimensions (Fig. 1d), or PyMC's graphs [7] where nodes contain the name of the prototype distribution of the variables (Fig. 1e). The Kruschke-style diagram [8] (Fig. 1f) elaborates the graph with the iconic "prototypes" of the variables' distribution on each node and annotations for the parameters of distributions being set by variables in the model.

Given a graph, users could view relations among variables (L1) at a glance (through the existence or absence of edges). In the case of the more informed graphs like Kruschke diagrams, users could even observe the exact statistical associations or mathematical equations (L3). But inferring the polarity of the effect of a variable on other variables (L2) is still very much dependent on the ability of the users to understand the mathematical details.

3.2 Visualizations of Variables' Relations

To convey relations' polarity (L2) visually, we need to incorporate representations of the model's real-data uncertainty.

Joint & Marginal Distributions. A model's joint distribution is multivariate. We could represent the pairwise joint or marginal distributions of the variables (Fig. 2a-c). While KDE (Kernel Density Estimation) plots are a common way of representing marginal distributions, 3D surfaces are rarely used for representing the pairwise joint distributions especially in the context of probabilistic modeling. Contour and scatter plots are more commonly used for this instead. There are various existing visualization libraries to create such representations for Bayesian analysis (ArviZ [25], bayesplot [26], tidybayes [27], shinystan [28]).

Variables' relations (L1), their polarity (L2) and aspects of their quantification (e.g. statistical associations) (L3) are conveyed by the shape of scatter and contour plots. Conditioning could help interpreting the shapes of these plots in regards with these three aspects of variables' relations, and retrieving them from KDE plots alone, as we explain below.

Scatter Plots. The shape of a scatter plot of samples and contours representing a 2D distribution can reveal relations between the two variables. For example, the well-elongated elliptical shape of the scatter plot of `temperature` and `b` in Fig. 2d implies the existence of a relation, while the rectangular shape of the scatter plot of `temperature` and `a` the absence of a relation (L1). In the first case, increasing values of `b` lead to higher mean value of the distribution of `temperature`, while in the later, increasing values of `a`

do not affect the distribution of `temperature`. The shape of the scatter plot reveals the polarity of the relation (L2) and the statistical associations (`b` controls the μ parameter of `temperature`'s distribution) (L3). The effect becomes more evident if we divide the sample set into subsets of samples for sequential increasing ranges of `b` (Fig. 2e-g).

Interactive Conditioning with IPME. Interactive conditioning has also been used to convey information about variables' relations through KDE plots. Taka et al. [9] suggest the interactive conditioning of the marginal distributions and the presentation of the *conditional marginal distributions* of the variables in their IPME tool. For example, comparing the marginal distribution of `temperature` (drawn in blue) with its three sequential conditional marginal distributions (drawn in orange) in Fig. 2m-o while conditioning on `b` in three increasing and sequential ranges, we could infer a relation such that increasing values of `b` lead to higher mean value of the distribution of `temperature` (reveals L1, L2, and statistical associations in L3). Conditioning is applied by the user by dragging a fixed-height and variable-width selection box in KDE plot corresponding to the conditioning variable. A video demonstrating IPME's interactivity can be found in [29].

Scatter Plot Matrix. A Scatter plot matrix (or pair plot) presents the pairwise joint and marginal distributions of a model's variables. ArviZ offers the ArviZ Point Estimate Pairplot (APEP) [30], which presents variables' joint samples and contours of the pairwise distributions on the bottom corner of the matrix and the KDE plots of the marginal distributions on the diagonal. Scatter plot matrices usually offer selection tools for applying data filtering (conditioning). We introduce in this paper the Interactive Pair Plot (IPP), an interactive scatter plot matrix like APEP that incorporates IPME's [9] interactive conditioning on the KDE plots to present the conditional marginal distributions. We present IPP in more detail in Section 4.

3.3 Evaluation of Visualization in Bayesian Reasoning

The effect of problems' representations on people's ability to reason about difficult and unintutive mathematical concepts has been investigated in the existing literature. A characteristic example is Bayesian reasoning where people seem to perform poorly when they have to update their believes in the light of new data (apply Bayes' rule) [19].

People's performance in Bayesian reasoning seemed to have been benefited when graphical displays (contingency tables, signal detection bar, detection bar, probability map or double-tree diagram) [19, 31] or iconic pictorial representations [18] or interactive frequency grids with check boxes [17] were combined with a textual description of the Bayesian reasoning problem. Expanding the sample through crowd-sourcing [15, 16] led to inconsistent findings with that previous work possibly because the wording of textual descriptions could significantly impact users' accuracy [14]. Ottley et al. [14] showed that (text-only or) visualization-only designs were more effective than those which blend text and visualization.

Interaction is believed to enable the communication between users and visual systems and support cognitive processing. However, it is not clear how beneficial interaction could be when added to a static representation. There are few studies having investigated this effect on users' performance in contexts like Bayesian reasoning, and the findings were unexpected. Mosca et al. [13] found no improvement in people's Bayesian reasoning by adding interactivity to static icon arrays through check boxes. Khan et al. [31] found that adding interactivity to double tree diagrams through dragging and dropping to increase users' active engagement significantly decreased users' performance in Bayesian reasoning. Khan et al. [31] suggest that people's worse performance when using interaction might result from the cognitive overload caused to them by interacting. The existing work about the added value of interaction on static visualizations is little and thus, conclusions about this contradiction cannot be easily drawn.

Brushing-and-linking is an interactive approach usually used on static visualizations of multivariate data like scatter plot matrices. This method is useful in many tasks like analyzing subsets of multivariate [24, 32–34] or hierarchical [23] data, solving conditioning problems or conducting sensitivity analysis [9], which could not easily be conducted through static visualizations. The added value of brushing-and-linking to static visualizations is rarely the main focus of evaluation user studies. For example, Nguyen et al. [32] found that an interactive version of a scatter plot visualization improved the accuracy of users in data exploration compared to static versions of it. The effect of brushing-and-linking alone could not be evaluated through this study design because the provided interactivity was ranging from choosing the number of plot panels to brushing-and-linking.

We designed a user study to investigate whether users can infer structural information about probabilistic models presented in scatter plot matrices. We followed a visualization-only design as it seems more effective based on the existing literature; the questions were including the visualization and no textual or mathematical (e.g. probabilistic statements) description of the model. We investigated the effect of adding interactive conditioning through brushing-and-linking like in IPME [9] to the scatter plot matrix. We aimed to provide context for the added value of brushing-and-linking in the exploration of multidimensional spaces of uncertainty.

4 Interactive Pair Plot

We designed an interactive pair plot (IPP) (Fig. 3) by combining elements from the designs of APEP and IPME visualizations, and adding some extra highlighting. We replicated the design of the APEP for the outlook of the pair plot. The scatter and contour plots of variables' pairwise joint samples and distribution are presented on the columns and rows of the matrix's lower triangle, and the KDE plots of the variables' marginal distributions on the diagonal.

IPP was built on IPME's framework inheriting its design elements (e.g. plot's style and attributes like the grey background, color palettes, rug plots' inclusion of variables' samples on the KDE plots, side interactive toolbar, and the interactive conditioning's mechanism and design) and limitations (e.g. inflexibility in rerunning online (prior) sampling or inference to get more samples in sub-ranges of model's

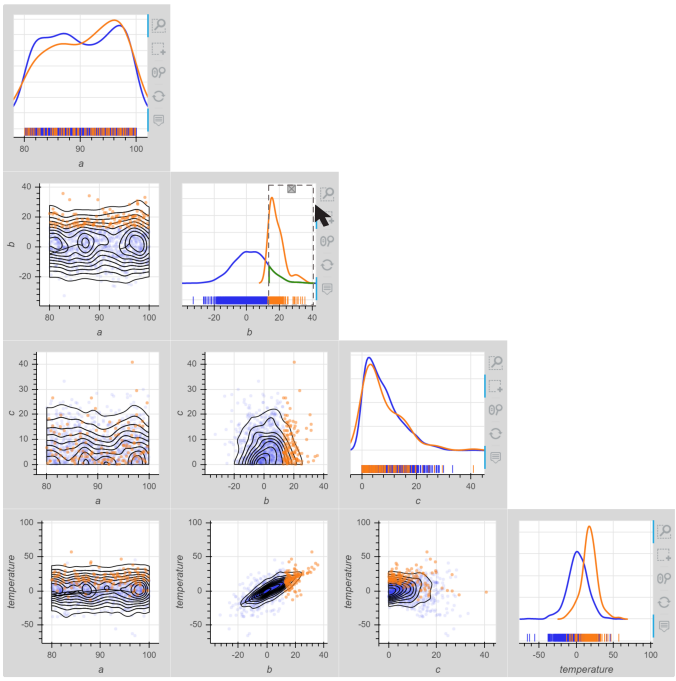


Fig. 3. IPP of Model's 1 variables. A selection box is dragged and drawn on the KDE plot of variable `b` restricting its range to $[12 - 20]$. The conditional marginal distributions of the variables are drawn (in orange) in the KDE plots on the diagonal and the samples in the restricted sample space are highlighted (in orange) in the scatter and rug plots.

sample space with few or no samples, or applying multiple conditions on a single variable).

The interactive conditioning's design in IPP replicates IPME's corresponding brushing-and-linking interactivity. The KDE plots (on the diagonal) can be interactively conditioned by dragging and drawing a fixed-height and variable-width selection box (brushing) and the KDE and rug plots are updated (linking) with the KDEs of the conditional distributions being drawn and samples in the restricted sample space being highlighted in orange exactly like in IPME. The choice of the selection box was limited by the offered options of the Bokeh visualization library used in the backend. We enhanced the interactive conditioning's design in IPP by adding the highlighting of the joint samples in the scatter plots (linking) in orange color for consistency. We aimed to replicate the typical linking effect encountered in scatter plot matrices. A video demonstrating IPP's interactivity can be found in [29].

We kept in IPP both the discrete (scatter plots, rug plots) and continuous (KDEs, contours) representations of model's distribution encountered in APEP and IPME to complement each other. The contours and KDEs illustrate how the density of the samples change with the value of the variables. Without these plots, identification of high probability density value ranges would be difficult especially in cases of high-density sampling that creates visual overlaps of samples. The scatter plots and rug plots of samples provide a discretized form of the continuous representations, which according to existing literature could better support people's reasoning for uncertainty [10, 11].

IPP's API considers subsets of variables of interest to deal with the quadratic scaling in area of the scatter plot matrix with the number of variables. This would be especially useful, for example, when users might want to inspect only an aspect of complex models with many variables or deep-hierarchy structures.

5 Research Questions, Tasks & Conditions

Does interactive conditioning when used on pair plots help users understand the structure of probabilistic models more accurately, faster, and with more confidence? We aimed to investigate how efficient interactive conditioning is in the comprehension of probabilistic models and when it can be beneficial. **Are there levels of detail of variables' relations or model designs for which interactive conditioning is beneficial?** We determined three types of tasks in the user study, T1, T2, and T3, each accounting for the L1, L2, and L3 level of detail, respectively. We explored various model designs as described in Section 6.1. Table 1 summarizes the study's tasks and models.

We designed a between-subject user study with two conditions; the static pair plot and the interactive pair plot (IPP). Participants in both conditions were viewing the same pair plot designed as described in the previous section, but participants in the static condition were not able to use the interactive conditioning. Fig. 4a presents a T2 task of the user study with a static pair plot shown to participants in the **static group (SG)** and an interactive pair plot (Fig. 4b) shown to participants in the **interaction group (IG)** instead.

Participants had to look at the static scatter plot matrices (in SG), or interact with them and look at the additional highlights (in IG) to perform the tasks. For example, participants in SG could determine the direction of relation between `b` and `temperature` in task `t3` (Fig. 4a) based on the shape of the scatter plot. Participants in IG could use interactive conditioning on increasing ranges of `b` and observe the changes in the highlighted visual elements of variable `temperature`. The fourth training video used in the user study (more about training in Section 6) presents a demonstration of task examples and how they could be answered by each group based on the presented visualization.

This design of the pair plot allows a fair comparison of the two conditions in regards with the amount of presented information. The changes in the interactive condition refer to two new types of visual elements added to the static visualization; firstly, a highlighting of visual information already existing in the static case (of dots in the scatter plots or vertical lines in the rug plots), and secondly the inclusion of the conditional marginal distributions in the KDE plots. These visual additions are informationally equivalent to scatter plots as the first does not insert new information, and the second represents existing information (in scatter plots' shape) in a different format (marginal distribution).

6 User Study's Design

The study was approved by the institution's ethics review board (approval number: 300200319) and conducted online. It consisted of three parts; **training**, **tasks**, and **demographic questions**. The training consisted of 4 videos (find links in supplemental material) presenting the aim and structure of study, an introduction to basic probabilistic concepts, an explanation of assigned version (static or interactive) of visualization, and some example tasks.

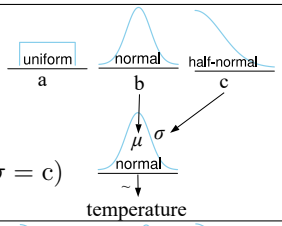
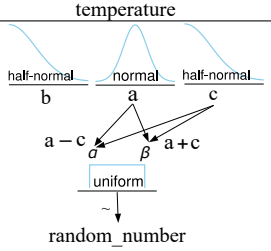
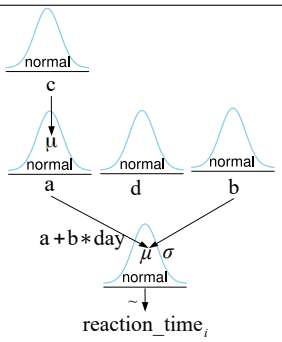
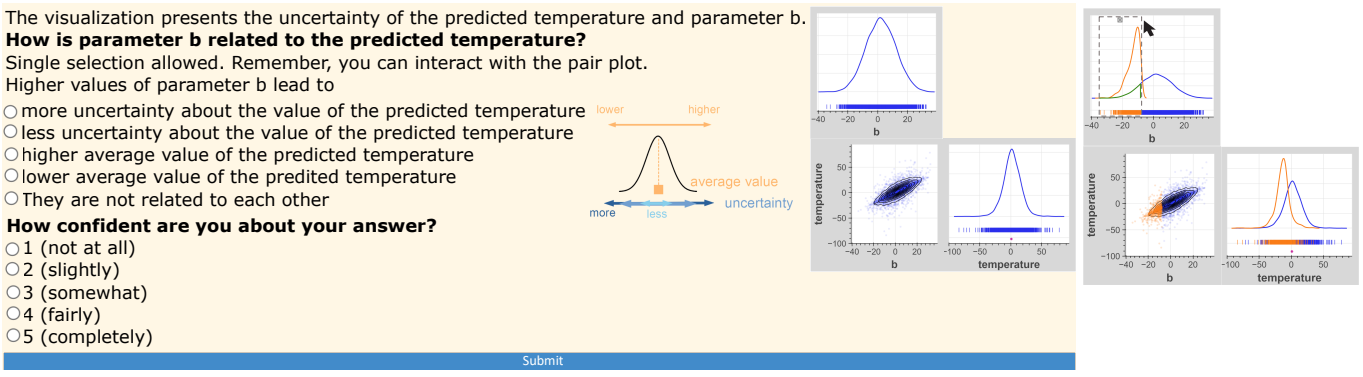
Model	Graph	Task id	T	Question
Model 1 $a \sim \text{Uniform}(\alpha = 80, \beta = 100$ $b \sim \text{Normal}(\mu = 2, \sigma = 10)$ $c \sim \text{Half-Normal}(\sigma = 10)$ **temperature** $\sim \text{Normal}(\mu = b, \sigma = c)$		t1	T1	Which of the parameters a, b and c are related to temperature?
		t2	T2	How is parameter a related to temperature?
		t3	T2	How is parameter b related to temperature?
		t4	T2	How is parameter c related to temperature?
		t5	T3	How would you describe the effect of parameters a, b and c on temperature?
Model 2 $a \sim \text{Normal}(\mu = 0, \sigma = 10)$ $b \sim \text{Half-Normal}(\sigma = 10)$ $c \sim \text{Half-Normal}(\sigma = 20)$ **random_number** $\sim \text{Uniform}($ $\alpha = a - c,$ $\beta = a + c)$		t6	T1	Which of the parameters a, b and c are related to random_number?
		t7	T2	How is parameter a related to random_number?
		t8	T2	How is parameter b related to random_number?
		t9	T2	How is parameter c related to random_number?
		t10	T3	How would you describe the effect of parameters a, b and c on lower_bound?
		t11	T3	How would you describe the effect of parameters a, b and c on upper_bound?
Model 3 $c \sim \text{Normal}(\mu = 100, \sigma = 150)$ $e \sim \text{Half-Normal}(\sigma = 150)$ $f \sim \text{Normal}(\mu = 10, \sigma = 100)$ $g \sim \text{Half-Normal}(\sigma = 100)$ $h \sim \text{Half-Normal}(\sigma = 200)$ $a_i \sim \text{Normal}(\mu = c, \sigma = e)$ $b_i \sim \text{Normal}(\mu = f, \sigma = g)$ $\text{sigma}_i \sim \text{Half-Normal}(\sigma = h)$ $d \sim \text{Normal}(\mu = 0, \sigma = 10)$ **reaction_time_i** $\sim \text{Normal}($ $\mu = a_i + \textbf{day} \cdot b_i,$ $\sigma = \text{sigma}_i)$		t12	T1	Which of the parameters a, b, c and d are related to reaction_time?
		t13	T1	Which of the parameters b, c and d are related to a?
		t14	T2	How is parameter a related to reaction_time?
		t15	T2	How is parameter b related to reaction_time?
		t16	T2	How is parameter c related to reaction_time?
		t17	T2	How is parameter d related to reaction_time?
		t18	T3	If reaction_time, a and c lie on a graph, what is the structure of the graph?
		t19	T3	How would you describe the effect of parameters a, b and day on reaction_time?

TABLE 1
Summary of probabilistic models and tasks used in the user study. The models' definitions and graphs are presented in the first two columns and the task id, task type (T), and question asked in the rest columns in the order presented to participants.



(a) (b)

Fig. 4. (a) Task t3 (Model 1 - T2) of user study. Participants in SG were shown a static pair plot. (b) The interactive pair plot participants in IG were shown instead. Both pair plots showed the minimum necessary subset of model's variables.

The study tasks were split into three parts, each corresponding to a different probabilistic model of increasing

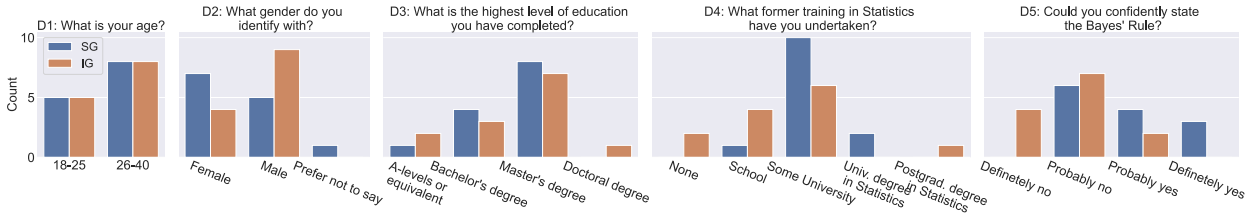


Fig. 5. **Demographic statistics** of participants in the user study. Both groups of participants (SG and IG) comprised of more older participants (D1). There was a slight gender imbalance between the groups with IG having more males and SG more females (D2). The educational background was generally well-balanced between the groups (D3), while participants in SG had a slightly higher former training in Statistics (D4, D5).

complexity. A set of questions of all three task types (Ts) was created for each model. Table 1 presents a summary of the 3 models and 19 tasks used in the study. The supplemental material provides a detailed description of the models and screenshots of all tasks. The models were presented in increasing complexity and the tasks in increasing level of structural detail, and in the same order to all participants (from `t1` to `t19`). The version of the scatter plot matrix was only varying among participants.

All questions were multiple-choice. Multiple selections were allowed for the T1 tasks, and single selection for the rest. Each available option was graphically illustrated in the cases of T2 and T3 questions. Participants' confidence was input in a five level Likert scale.

At the end participants recorded their age, gender, highest educational level completed, former training in Statistics and knowledge of Bayes' rule. Fig. 5 presents participants' demographic statistics. We randomly assigned 26 participants in the IG and SG (13 participants in each). Participants were recruited through mailing lists and social media of the institution and personal contacts without any requirement regarding their statistical background. They each received a £10 worth online shopping voucher as a compensation.

6.1 Task Models' Design

We aimed to include different models of increasing complexity in the study. That was achieved by increasing the number of variables used for setting parameters of the observed variable's distribution in each model, combining various mathematical operations (+,-,*) for the assignment of distributions' parameters, and the use of hierarchy (Model 3). We aimed to include both typical (Model 1 and 3) and more exotic (Model 2) model designs to account for any possible prior familiarity of participants with certain statistical structures, and the use of a variety of distribution types.

Model 1 was the simplest probabilistic model used in the user study and is a typical one; a normal distribution for the observed variable with the mean and variance being directly set by two other (latent) parameters of the model.

Model 2 used a parameterization to set the bounds of the observed variable's uniform distribution through a deterministic transformation: $\alpha = \mathtt{a} - \mathtt{c}$ and $\beta = \mathtt{a} + \mathtt{c}$. This parameterization broadens the visual effects we can explore. The combination of a uniform (`temperature`) and normal (`a`) or half-normal (`c`) distribution creates unusual shapes of the pairwise scatter plots. The interpretation of the changes in the conditional marginal distribution while interacting is different in this model in comparison to previous model, because it is the bounds of the distribution that change here.

Model 3 was the most complex model representing a typical hierarchical linear regression model with a normal distribution for the observed variable, an often encountered structure in probabilistic modeling. The mean of the distribution was set as $\mu = \mathtt{a} + \mathtt{b} * \mathtt{day}$ and there were hyper-priors set for the priors of the `a` and `b` parameters. The hierarchy of the latent parameters in this model is an added complexity in comparison to previous models.

The observed random (or deterministic) variable in each model had a semantically meaningful name (`temperature`, `random_number`, `reaction_time`, `day` (deterministic)). The unidentified parameters were named with letters `a`,`b`,`c` etc. to avoid revealing information about variables' relations through their names (e.g. sigma, mu). Each model had an unidentified parameter which was *unrelated* to the rest of variables. We used a variety of prior distributions for the unrelated parameters; a uniform for parameter `a` in Model 1, a half-normal for parameter `b` in Model 2, and a normal for parameter `d` in Model 3.

Models' prior distributions were used in the study. Prior distributions reflect directly models' structure. As observations come into a model and the prior beliefs are updated, the initial structure of the model can be overwhelmed in the posterior distribution. For a clearer experimental protocol, we focused on the *prior space*.

6.2 Implementation Details

Irrelevant interactive elements (zoom tools, hovering-over tooltips, tabs, drop-down menus) were removed from pair plots in study tasks to isolate the conditioning-related interactivity as the focus of the study was on that. The pair plot was showing the minimum necessary subset of model's variables in every task to avoid overwhelming participants with irrelevant information. The unidentified variables were appearing in alphabetical order across the diagonal of the matrix with the observed variable presented at the bottom to create a consistent view across participants and tasks and avoid any possible extra cognitive load of participants having to search for a variable in a randomized matrix.

The task models were specified and interpreted in PyMC3. PyMC3's prior sampling (`pymc3.sample_prior_predictive`) was used to generate prior samples for models' variables. For example, for Model 1 specified in PyMC3 in Fig. 1b, we generated samples from the prior joint distribution of `temperature` and `b` represented by the blue surface in Fig. 2a. We used this set of samples to create the scatter and KDE plots in Fig. 2d-g and l-o, Fig. 3, and Fig. 4a,b.

7 Analysis and Results

7.1 Evaluation Measures

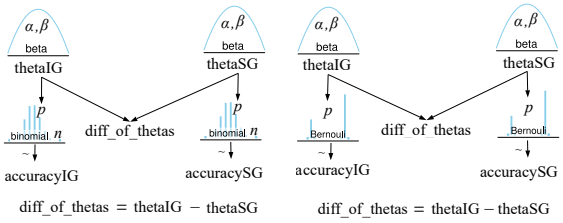
Participants' accuracy, response time and confidence were evaluated in the user study. Accuracy was measured as the number of correct selections in the multiple-choice input by each participant in every task. Participants' selections in the multiple-choice input were transformed into a binary representation with 0 indicating a wrong and 1 a correct selection. The binary representation of each response in T1's tasks (multiple selections were allowed) consisted of as many binary digits as the available options of the multiple-choice input, excluding the "none" option, while for T2 and T3 tasks (single selection was allowed) consisted of a single digit. Participants' performance in each task was computed as the number of occurrences of digit 1 in their response.

Participants' response time was measured (in seconds) from the moment the task was displayed until the answer was submitted. Participants were rating their confidence in each task on a 1-5 scale with increasing level of confidence (1: not at all, 2: slightly, 3: somewhat, 4: fairly, 5: completely). We remapped this to the $\{-2, -1, 0, 1, 2\}$ set to center the parameterization.

7.2 Data & Bayesian Modeling of Responses

Data. The data was split into sub-sets based on the condition (IG and SG). No participant or response was excluded. We did not consider the multiple blank registrations of some participants, who accidentally clicked the "Register" button multiple times. Accuracy data consisted of numbers of participants' correct selections in the multiple-choice input in every task. Response time data consisted of times (in sec). Confidence data consisted of ordinal values. A Bayesian analysis of the collected data was conducted on the level of the individual tasks.

Accuracy Analysis Models. Each group's performance in every task was modelled by a binomial likelihood (that was reduced to a bernoulli likelihood for T2 and T3 tasks). The posterior *probability of success* θ of the binomial distribution was estimated for each group. This probability expresses the propensity of a participant in the corresponding group to make a correct selection in each task. The two groups were compared in terms of accuracy by taking the difference of each group's posterior distribution of θ.



(a) **Accuracy** in T1 tasks (b) **Accuracy** in T2-T3 tasks

Fig. 6. Kruschke-style diagrams of the **accuracy analysis models**.

Response Time Analysis Model. Each group's response time in every task was modelled by a normal likelihood. The posterior distribution of *effect size (Cohen's d)* was estimated for the comparison of the two groups to normalise for the varying duration (and thus typical variances) of the tasks.

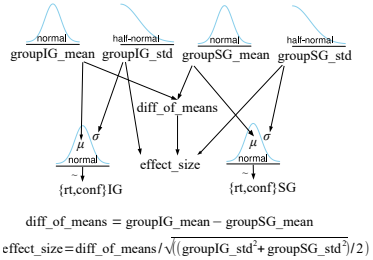


Fig. 7. Kruschke-style diagrams of the **response times and confidence analysis models**.

Confidence Analysis Model. Each group's confidence in every task was modelled by a normal likelihood. We made the simplifying assumption that the ordinal values could be treated as if they lay on a common continuous scale; hence the normal likelihood. A more sophisticated analysis could have inferred a (potentially per-subject) monotonic relationship between ordinal responses and "true" confidence. The posterior *mean confidence level* was estimated for each group as confidence takes ordinal values and there was no need to normalise. The difference of the mean confidence posterior distribution of each group for every task were estimated to compare the two groups.

7.3 Results

Fig. 8a presents the results of inference in a set of forest plots. Comparing the two groups (IG and SG) based on the differences of the posterior distributions, an effect of the interactive conditioning is more likely given the data as the value 0.0 (reference line in columns 3-5 of Fig. 8 indicating no difference) becomes less likely under the difference of the posterior distributions (horizontal posterior highest density interval bars). That's the highest density intervals of the posteriors in the forest plots presenting the differences are pulled away from the reference value towards the right. The effect of interactive conditioning becomes less likely when the highest density intervals of the posteriors are pulled towards the reference value.

Accuracy. Participants' performance overall was good in both groups with the inferred probability θ of giving a correct answer being over 0.5 in most tasks with greater certainty for tasks of lower level of structural detail (T1-T2) (columns **1-2** of Fig. 8a). Participants' accuracy in tasks `t3-4`, `t11`, `t19` in IG and `t19` in SG was around 0.5, while in tasks `t9-10` in IG and `t7-11`, `t18` in SG was the lowest than other tasks (< 0.5). The lower accuracy concerned tasks with instances of more complicated statistical modeling; the parameterization of the bounds of a uniform distribution in `t7-11` (Model 2) and a hierarchical structure of a hyper-prior and prior in `t18` (Model 3).

The effect of interactive conditioning is revealed by the differences of θs. The effect is strong in `t6-8` (Model 2), and `t18` (Model 3) and weaker in `t10-11` (Model 2) (column **3** of Fig. 8a) clearly showing the benefit of interactive conditioning in more sophisticated model designs.

Tasks `t2` (Model 1), `t8` (Model 2), and `t17` (Model 3) concerned unrelated variables. We observe a strong effect

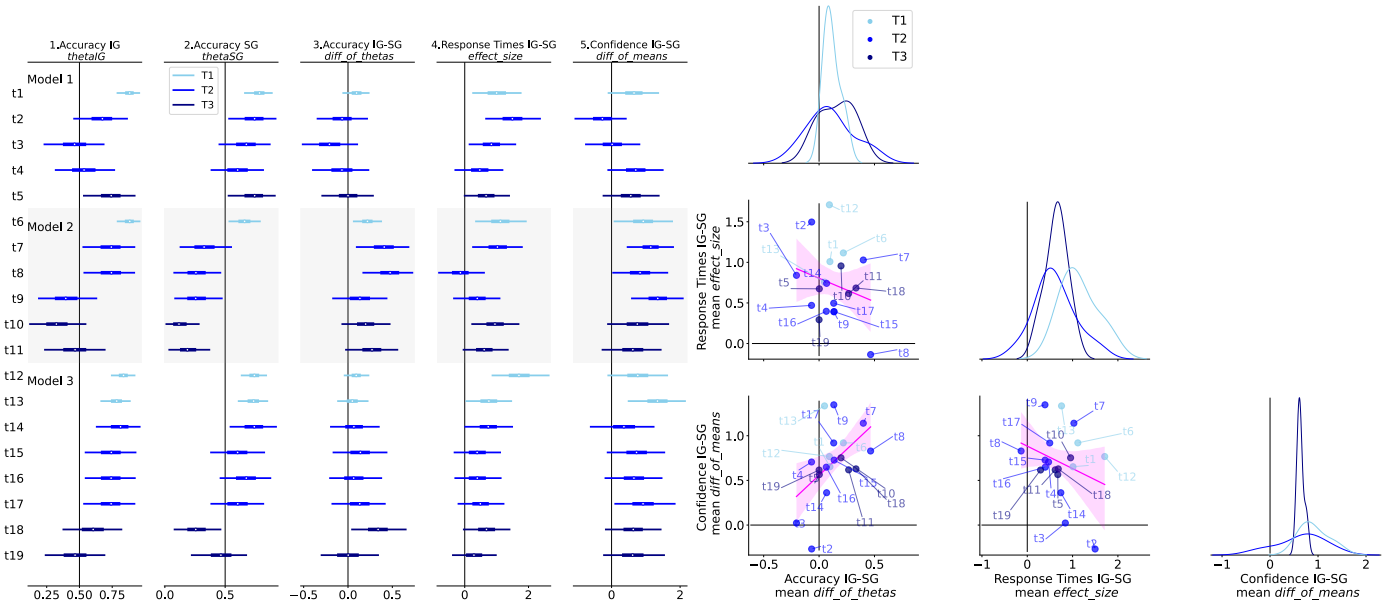


Fig. 8. **Results**. (a) Forest plot (94% highest density intervals) of the posterior distributions of the probability of correct answer for IG (*thetaIG*) and SG (*thetaSG*), difference of θs (*diff_of_thetas*), effect size of response times (*effect_size*) between IG and SG (normalised difference of duration), and difference of the estimated mean confidence of participants about their responses (*diff_of_means*). Tasks are presented vertically grouped per model. (b) Pair plot of mean values of the posterior distributions of *diff_of_thetas* for the accuracy, *effect_size* for the response times and *diff_of_means* for the confidence. The fitted linear regression line is drawn with a 90% bootstrap confidence interval in each scatter plot.

of interactive conditioning in `t8` (Model 2) and not in the rest (columns **3** of Fig. 8a). The square and full-gaussian shapes of the scatter plots in `t2` and `t17` respectively were more accurately interpreted as "absence of relation", while the half-gaussian shape of the scatter plot in `t8` confused participants in SG regarding the existence of a relation (see relevant screenshots of tasks in supplemental material). This shows the benefit of interactive conditioning in identifying relations in cases of peculiar shapes of scatter plots resulting from unusual combinations of prior distributions.

Response Times. Participants using interactive conditioning needed more time to complete tasks overall with the effect being strong in tasks `t1-3` (Model 1), `t6-7` and `t10` (Model 2), `t12-14` (Model 3) (column **4** of Fig. 8a). The differences in response time between groups are pulled towards the reference line (0.0) in tasks of middle or high level of detail `t4-5` (Model 1), `t8-9` and `t11` (Model 2), `t15-17` and `t18-19` (Model 3) (column **4** of Fig. 8a) implying that the extra exploration time interactive conditioning introduces tends to diminish in cases of more complex tasks.

Task `t8` was the one with the smallest mean difference (close to 0.0) in the response time and the greatest mean difference in accuracy between the groups (columns **3**-**4** of Fig. 8a). This could imply that the observed effect on accuracy cannot be explained by possible extra exploration time in IG (see Section 7.4 for further analysis on this).

Confidence. Participants in IG are more confident than those in SG overall with the effect being strong in tasks of lower level of detail `t6-9` (Model 2), `t13` and `t17` (Model 3) (column **5** in Fig. 8a). Task `t13` presents one of the strongest effects of interactive conditioning on confidence, while there is no corresponding effect on accuracy (columns **3**, **5** in Fig. 8a). This could imply that interactive conditioning makes participants with equivalent performance more confident (see Section 7.4 for further analysis on this).

7.4 Comparative Analysis

Do higher response times imply better accuracy or higher confidence? Does better accuracy imply higher confidence? The conduction of a causal analysis of the observed data is out of the scope of this analysis, but we will investigate the existence of relations (correlations) between these pairs based on the inferred data.

Fig. 8b presents the pair plot of the mean values of the differences of the posteriors for the accuracy, response times, and confidence between the IG and SG groups. There is a positive correlation between the differences in accuracy and confidence implying increased confidence with increased accuracy of the IG in comparison to the SG. Interactive conditioning when used in pair plots seems to support more accurate and certain decisions in the study tasks than the static pair plots. There is a slight negative correlation between the differences in accuracy and response time implying that any increase in the accuracy of the IG would not be attributed to increased response times. The negative correlation between the differences in confidence and response time is implied by the previous two correlations.

7.5 Analysis of Interaction Logs

The coordinates of the selection boxes drawn by the IG participants in each task were recorded. The proportion of participants having drawn at least one selection box was high in all tasks ($10/13$ in task `t5`, $12/13$ in tasks `t3-4`, `t6`, `t15`, `t19`, and $13/13$ in all other tasks). The (Q1,Q2,Q3) quartiles of the number of selection boxes drawn per task were (4.5, 9., 13.) and of the normalized (by the range of the corresponding variable) length of selection boxes were (0.11, 0.16, 0.24). The lengths of the selection boxes dragged and drawn by participants were between the 10% and 25% of variables' ranges being coherent with the shape of the

distributions. Such sizes of selection boxes are big enough to capture part of the distribution with roughly constant curvature. Bigger sizes would capture several modes.

8 Discussion

8.1 When is Interactive Conditioning (not) Beneficial?

Model Designs. The analysis of the collected data (Section 7.3) showed that the use of interactive conditioning is beneficial for users' comprehension of probabilistic models in cases of more sophisticated model designs like the parameterization of the bounds of a uniform distribution, hierarchical structures, and unusual combinations of prior distributions (e.g. uniform and half-normal) (`t6-8`, `t18`). This finding is strengthened by the fact that participants in SG had a quite higher former training in Statistics than those in IG (Fig. 5D4-5) and by excluding the possibility that the higher accuracy could be attributed to the longer response times observed in IG due to more time spent in the exploration of the structures (Section 7.4).

In all other tasks (`t1-5`, `t10-17`, `t19`) there was no strong effect of interactive conditioning on participants' performance (column **3** of Fig. 8a). We excluded the possibility of low use of interactive conditioning being an explanation for this (Section 7.5). The low complexity and commonness of the statistical associations encountered in most of these tasks or the high complexity of others could possibly, at least partly, explain this observation. Tasks `t1-5` and `t12-17` concerned simple common statistical associations (e.g. a normal distribution setting the μ of another normal distribution) making both representations adequate enough for participants to achieve a similarly good performance ($\theta > 0.5$) in most of these tasks (columns **1-2** in Fig. 8a). Task `t19` was a very complex task concerning the determination of the mathematical formula (a linear regression) (T3) linking four variables (`temperature`, `a`, `b`, `day`) together. Participants' performance in any of the groups was close the random choice ($\theta \approx 0.5$) (columns **1-2** in Fig. 8a).

Level of Structural Detail. Interactive conditioning did not seem to benefit participants' performance in a specific level of detail of variables' relations (column **3** of Fig. 8a). Model design was more determining in the effect of interactive conditioning on users' performance than the complexity of the task. On the contrary, the level of detail seem to play a role in the differences of participants' response time and confidence between the two groups. The differences in response time diminish in tasks of higher levels of detail (Section 7.3). The burden of the extra exploration time required for interactive conditioning reduces in more complex tasks. The greatest advantage of interactive conditioning in participants' confidence in comparison to static pair plots appear with greater certainty in tasks of lower level of detail (Section 7.3). In simpler tasks, participants using interactive conditioning are more confident.

8.2 Practical Implications

Various types of users of probabilistic models could benefit from visualizations like interactive pair plots. Model builders could benefit from such visualizations when they encounter complex model designs or have lack of statistical experience to conduct prior predictive checks and validate models (are the relations (effects) of variables as expected?). Decision-makers in crucial areas like healthcare or stock market could benefit from such visualizations when they should eliminate any risk of ignorance or misunderstanding of models' structure to decide on crucial interventions (what is the effect of a model's variable on another?).

Researchers could benefit from such visualizations when they need to tune some parameters in a model and need to (for)see the effect of doing so, or to communicate their models to a broader audience and provide a more intuitive overview of the model. Teachers and learners of Bayesian modeling could benefit from such visualizations when the former seek for ways to illustrate the effects of variables in various model designs, and the latter to gain a more intuitive understanding of the different model designs.

Visualizations like IPP could help users explore the effects of variables in the posterior space, as well. The relationships of variables under the posterior are governed by effects, although these effects usually cannot be expressed analytically in explicit mathematical or statistical associations as they can in the prior space. This happens because the exact form of the posterior distribution usually cannot be expressed analytically unless the priors are conjugate [35]. In these cases, posterior distributions can be estimated by sampling algorithms (e.g. MCMC), which does not allow us to know how exactly variables are associated (e.g. variable `x` sets the μ of the distribution of variable `y`). Although the effects of variables in the posterior space could be visualized and explored though such visualizations, most probably they could not be interpreted as specific analytical relations.

The findings of the user study are not restrictive to a specific PPL. Visualizations like IPP could be used to present the output of any probabilistic programming code that is being interpreted. Any programming language or library (including PPLs) that supports operations on probability distributions could be used for sampling from the prior and any PPL could be used for inferring the posterior.

8.3 Limitations of User Study

The analysis of the collected data suggests that interactive conditioning is beneficial for users' understanding, but it is not clear what aspect of it actually helps users. The recorded interaction data could not provide us with more insight into how participants in IG were exploiting the specific implementation of interactive conditioning. Were they combining information from both scatter plots and marginal distributions? Were they only looking at the conditional marginal distributions? Or were they only looking at the highlighted scatter plots? Such questions would require other experiment designs that would include one or combinations of open questions, think-aloud protocol, analysis of participants' micro-interactions [36] or eye-tracking.

The participants' sample in this study present limited demographics with respect to age and educational background. Further experimentation could be conducted on an expanded sample with broader demographics to investigate if the findings of this user study would replicate (as Ottley et al. [16] did for Brase [18] and Micallef et al. [15]).

We focused on presenting the study's models in the prior space. This offers analytic descriptions of variable's relations

(what controls what and how) as explained in Section 8.2, which could be used to validate participants' responses (ground truth). For a clearer experiment design, we did not include the case of posteriors determined by conjugate priors. The types of distributions we could explore was limited by the fact that prior sampling from heavy tail distributions (student-t, Pareto, Cauchy) gives a Dirac-delta-looking estimation of the probability density. Exploring variables distributed in such ways in IPP would not reveal any effect on their distribution while conditioning on them.

We had to limit the number of questions to ensure the completion of study by participants in roughly an hour. The user study was designed to include a variety of probabilistic model types (parameterized, linear regression, hierarchical), distributions (normal, half-normal, uniform), and statistical associations (setting the mean, standard deviation, or bounds of the distribution directly or through simple mathematical equations). There are many more model designs (logistic regression, GPs), distributions (discrete distributions like binomial and Poisson) and configurations that could be explored in the future in the context of a study like this one.

Variables' relations in a probabilistic model could also be characterized by their causal direction (directionality of the arrow that links two variables in the model's graph). This was not included in the relations' level-of-detail hierarchy described in Section 2.2 as a separate level to limit the task types, and hence questions in the user study. Participants' performance in T2 and T3 tasks could be used as an indication of whether they were able to infer this information. Inferring which variable controls a parameter of the distribution of another variable implies the inference of the causal direction between the variables.

8.4 Conclusions

Although there are various existing visualizations of probabilistic models and variables' relations, it is very little known about whether and when they support users' comprehension of the models. We focused on interactive conditioning and investigated through a user study whether adding it to classical scatter plot matrices helps users better understand probabilistic models and if there are levels of structural detail and model designs for which it is beneficial. The analysis of the collected data showed that interactive conditioning is beneficial in cases of sophisticated model designs and the difference in response time between the interaction and static group becomes less important in higher levels of structural detail. Participants using interactive conditioning were more confident about their responses overall with the effect being stronger in tasks of lower level of detail. We believe that these initial findings evoke the need for more research to understand how users can benefit from visual representations of probabilistic models and could pave the way for future investigation into the role of interaction to support more explainable Bayesian probabilistic models [37] and users' engagement with them.

ACKNOWLEDGMENTS

This work was supported by the Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics, EPSRC Project: EP/R018634/1. The data and analysis code can be found in [38].

REFERENCES

[1] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 903–913, 2019.

[2] M. Kay, G. L. Nelson, and E. B. Hekler, "Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI," in *Proc. CHI '16*, p. 4521–4532.

[3] C. Phelan, J. Hullman, M. Kay, and P. Resnick, "Some Prior(s) Experience Necessary: Templates for Getting Started With Bayesian Analysis," in *Proc. CHI '19*, p. 1–12.

[4] A. Sarma and M. Kay, "Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis," in *Proc. CHI '20*, p. 1–12.

[5] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Cambridge, MA, USA: The MIT Press, 2009.

[6] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn, *WinBUGS Version 2.0 Users Manual.*, 2003. [Online]. Available: https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf

[7] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull, "Graphviz and dynagraph – static and dynamic graph drawing tools," in *GRAPH DRAWING SOFTWARE*. Springer-Verlag, 2003, pp. 127–148.

[8] J. Kruschke, "Chapter 8: JAGS," in *Doing Bayesian Data Analysis (Second Edition)*. Boston: Academic Press, 2015, pp. 193–219.

[9] E. Taka, S. Stein, and J. H. Williamson, "Increasing interpretability of Bayesian probabilistic programming models through interactive representations," *Front. Comput. Sci.*, vol. 2, p. 52, 2020.

[10] J. Hullman, P. Resnick, and E. Adar, "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering," *PLOS ONE*, vol. 10, no. 11, pp. 1–25, 11 2015.

[11] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "When (Ish) is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems," in *Proc. CHI '16*, p. 5092–5103.

[12] M. Correll and M. Gleicher, "Error bars considered harmful: Exploring alternate encodings for mean and error," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2142–2151, 2014.

[13] A. Mosca, A. Ottley, and R. Chang, "Does Interaction Improve Bayesian Reasoning with Visualization?" in *Proc. CHI '21*.

[14] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. J. Han, and R. Chang, "Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 529–538, 2016.

[15] L. Micallef, P. Dragicevic, and J. Fekete, "Assessing the effect of visualizations on bayesian reasoning through crowdsourcing," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2536–2545, 2012.
[16] A. Ottley, B. Metevier, P. K. J. Han, and R. Chang, "Visually Communicating Bayesian Statistics to Laypersons," Tufts University, Tech. Rep., 2012.
[17] J. Tsai, S. Miller, and A. Kirlik, "Interactive Visualizations to Improve Bayesian Reasoning," *Proc. Hum. Factors Ergonom. Society Annu. Meeting*, vol. 55, pp. 385–389, 09 2011.
[18] G. L. Brase, "Pictorial representations in statistical reasoning," *Applied Cognitive Psychology*, vol. 23, no. 3, pp. 369–381, 2009.
[19] W. G. Cole, "Understanding Bayesian Reasoning via Graphical Displays," in *Proc. CHI '89*, p. 381–386.
[20] J. A. McDonald, "INTERACTIVE GRAPHICS FOR DATA ANALYSIS," Ph.D. dissertation, August 1982.
[21] C. M. Newton, "Graphics: From alpha to omega in data analysis," in *Graphical Representation of Multivariate Data*. Academic Press, 1978, pp. 59–92.
[22] R. A. Becker and W. S. Cleveland, "Brushing Scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, 1987.
[23] K. Sankaran and S. W. Holmes, "Interactive Visualization of Hierarchically Structured Data," *J. Comput. Graph. Stat.*, vol. 27 3, pp. 553–563, 2018.
[24] A. R. Martin and M. O. Ward, "High Dimensional Brushing for Interactive Exploration of Multivariate Data," in *Proc. VIS '95*, p. 271.
[25] R. Kumar, C. Carroll, A. Hartikainen, and O. A. Martin, "ArviZ a unified library for exploratory analysis of Bayesian models in Python," *The Journal of Open Source Software*, 2019.
[26] J. Gabry and T. Mahr, "bayesplot." [Online]. Available: https://mc-stan.org/bayesplot/
[27] M. Kay, *tidybayes*. [Online]. Available: http://mjskay.github.io/tidybayes/
[28] Stan Develop. Team, "shinystan." [Online]. Available: http://mc-stan.org/shinystan/
[29] Interactive Probabilistic Models Explorer. [Online]. Available: https://github.com/evdoxiataka/ipme
[30] ArviZ Point Estimate Pairplot. [Online]. Available: https://arviz-devs.github.io/arviz/examples/plot_pair_point_estimate.html
[31] A. Khan, S. Breslav, and K. Hornbæk, "Interactive Instruction in Bayesian Inference," *Human–Computer Interaction*, vol. 33, no. 3, pp. 207–233, 2018.
[32] Q. V. Nguyen, N. Miller, D. Arness, W. Huang, M. L. Huang, and S. Simoff, "Evaluation on interactive visualization data with scatterplots," *Visual Informatics*, vol. 4, no. 4, pp. 1–10, 2020.
[33] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1539–1148, 2008.
[34] Q. V. Nguyen, S. Simoff, Y. Qian, and M. L. Huang, "Deep exploration of multidimensional data with linkable scatterplots," in *Proc. VINCI '16*, p. 43–50.
[35] B. Lambert, "12.3. The Difficulty With Real-Life Bayesian Inference," in *A Student's Guide to Bayesian Statistics*. SAGE Publications, 2018, pp. 265–266.
[36] S. Breslav, A. Khan, and K. Hornbæk, "Mimic: Visual Analytics of Online Micro-Interactions," in *Proc. AVI '14*, p. 245–252.
[37] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," in *Proc. of IUI '15*, p. 126–137.
[38] E. Taka, S. Stein, and J. H. Williamson. (2022) Does interacting help users better understand the structure of probabilistic models? University of Glasgow Enlighten Repository. [Online]. Available: http://dx.doi.org/10.5525/gla.researchdata.1248

Evdoxia Taka received the Dipl.Ing. degree in Electrical and Computer Engineering and M.Sc. degree in Computer Science from Aristotle University of Thessaloniki, Greece, in 2014 and 2016, respectively. She is currently a Ph.D. candidate at the School of Computing Science, University of Glasgow, UK. Her main research interests include visualization, probabilistic modeling and programming, and causality.

Sebastian Stein received his Dipl.-Inf. from the Technical University of Dortmund, Germany, in 2010, and his PhD in Computing from the University of Dundee, UK, in 2014. His research interests are in intelligent interactive systems, spanning areas of HCI, ubiquitous computing, machine learning and probabilistic modelling.

John H. Williamson John H. Williamson is a Senior Lecturer at the University of Glasgow, where he received his Ph.D. and B.Sc. He has a long-standing interest in the use of machine learning and probabilistic modelling in human-computer interaction, and in Bayesian models of interaction.