




Blang: Bayesian Declarative Modeling of General Data Structures and Inference via Algorithms Based on Distribution Continua

Alexandre Bouchard-Côté 
University of British Columbia


Kevin Chern 
University of British Columbia

Davor Cubranic 
University of
British Columbia

Sahand Hosseini
University of
British Columbia

Justin Hume
Third Foundation Labs Inc.

Matteo Lepur
University of
British Columbia

Zihui Ouyang 
University of
British Columbia

Giorgio Sgarbi
University of
British Columbia

Abstract

Consider a Bayesian inference problem where a variable of interest does not take values in a Euclidean space. These “non-standard” data structures are in reality fairly common. They are frequently used in problems involving latent discrete factor models, networks, and domain specific problems such as sequence alignments and reconstructions, pedigrees, and phylogenies. In principle, Bayesian inference should be particularly well-suited in such scenarios, as the Bayesian paradigm provides a principled way to obtain confidence assessment for random variables of any type. However, much of the recent work on making Bayesian analysis more accessible and computationally efficient has focused on inference in Euclidean spaces.

In this paper, we introduce **Blang**, a domain specific language and library aimed at bridging this gap. **Blang** allows users to perform Bayesian analysis on arbitrary data types while using a declarative syntax similar to the popular family of probabilistic programming languages, **BUGS**. **Blang** is augmented with intuitive language additions to create data types of the user’s choosing. To perform inference at scale on such arbitrary state spaces, **Blang** leverages recent advances in sequential Monte Carlo and non-reversible Markov chain Monte Carlo methods.

Keywords: Bayesian modeling language, Bayesian inference, non-standard data structures, Blang.

1. Introduction

Blang is a probabilistic programming language (PPL) and software development kit (SDK) for performing Bayesian data analysis. Its design supports scalable inference over arbitrary data types, in particular, combinatorial spaces which are of central importance in areas such as computational biology. The design philosophy is centered around the day-to-day requirements of real-world data science. In the following, we put Blang in the context of the rich PPL and Bayesian modeling ecosystem.

Probabilistic programming has revolutionized applied Bayesian statistics in the past two decades; now being a part of the core toolbox of applied statistics. For example, packages such as BUGS (Lunn *et al.* 2000, 2009, 2012), JAGS (Plummer 2003), Stan (Carpenter *et al.* 2017), and PyMC3 (Salvatier, Wiecki, and Fonnesbeck 2016) have been widely used in various applications ranging from ecology (Semmens, Ward, Moore, and Darimont 2009), to astronomy (Greiner, Burgess, Savchenko, and Yu 2016), and psychology (Bürkner and Vuorre 2019). See Van de Meent, Paige, Yang, and Wood (2018) for a recent survey.

In recent years, research in the area of Bayesian modeling software has focused on two main directions. On one hand, considerable progress has been made in designing general-purpose PPLs (Wood, Van de Meent, and Mansinghka 2014; Paige and Wood 2014; Milch, Marthi, Russell, Sontag, Ong, and Kolobov 2005; Goodman, Mansinghka, Roy, Bonawitz, and Tenenbaum 2008) which are able to represent any computable probability distributions (Ackerman, Freer, and Roy 2017). However, inference in these powerful languages often has to resort to algorithms such as non-Markovian sequential Monte Carlo that can have poor scalability. Rapid progress is being made to lift this limitation (e.g., Paige and Wood 2016; Zhou, Gram-Hansen, Kohn, Rainforth, Yang, and Wood 2019; Ronquist *et al.* 2020) but in typical applications that involve challenging combinatorial spaces, general purpose PPL inference engines are not yet able to match the performance of specialized samplers. A second area of active development (Carpenter *et al.* 2017; Salvatier *et al.* 2016; Bingham *et al.* 2018, *inter alia*) has been to use automatic differentiation combined with Hamiltonian Monte Carlo (HMC) sampling (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 2011), which is highly efficient in problems defined on continuous state spaces. Naturally, algorithms based on HMC are not necessarily well-suited for inference problems defined on discrete and combinatorial state spaces.

In the past, efficient sampling in combinatorial spaces has been achieved by designing portfolios of specialized samplers in a case-by-case basis (see, e.g., Lakner, Van der Mark, Huelsenbeck, Larget, and Ronquist 2008). This process is typically time consuming and error prone. There is an opportunity to simplify this process, minimize manual intervention of tuning algorithms, and to speed-up and parallelize inference. This is possible with new developments in computational statistics such as non-reversible Markov chain Monte Carlo (MCMC) methods (Syed, Bouchard-Côté, Deligiannidis, and Doucet 2019) based on parallel tempering (TP; Geyer 1991), and a non-standard flavor of the sequential Monte Carlo (SMC) method that we call sequential change of measure (SCM, to avoid confusion with state-space SMC; Del Moral, Doucet, and Jasra 2006; Neal 2001) augmented with adaptive schemes (Zhou, Johansen, and Aston 2016). All these schemes are based on a continuum of probability distributions, all defined on the same space and interpolating between the prior and posterior. The benefit of these methods is that a simplistic set of sampling algorithms can still achieve high sampling efficiency while exploiting parallel architectures. Blang fully automates the construction of in-

terpolating probability distributions and therefore democratizes the use of high-performance Monte Carlo schemes such as non-reversible PT and SCM.

Blang is designed to be efficient not only in computational terms but also for the user's development time. To achieve this goal, considerable effort has been put to facilitate model construction, testing, reuse and integration into existing data analysis pipelines, and to support reproducible data analysis. Instead of creating a language from scratch, Blang is built using Xtext (Efftinge and Völter 2006), a powerful framework for designing programming languages. Owing to this infrastructure, Blang incorporates a feature set comparable to many modern, fully-fledged, multi-paradigm languages: functional, generic and object programming, static typing, just-in-time compilation, garbage collection, IDE support for static types, profiling, code coverage, and debugging.

Blang comes with a growing library of built-in models, which are themselves written in Blang (as done in Murray and Schön 2018), moreover, users can share and maintain models via an established transitive dependency management and versioning system. Blang also implements a suite of existing and novel testing strategies for models and MCMC methods, blending them with unit testing and multiple testing tools.

One of the existing PPLs most closely related to Blang is RevBayes (Höhna, Heath, Boussau, Landis, Ronquist, and Huelsenbeck 2014; Höhna *et al.* 2016). RevBayes is a declarative PPL which provides extensive support for Bayesian inference over phylogenetic trees, an archetypical example of a challenging combinatorial space. Moreover, RevBayes supports interactive usage, a functionality currently not supported in Blang. However, as phylogenetic inference is the primary domain targeted by RevBayes, users interested in combinatorial spaces other than phylogenetic trees will benefit from Blang's abstractions which target arbitrary combinatorial spaces.

The goal of this paper is to provide readers with an introduction to Blang. We begin with an outline of the language's goals in Section 2, describe the open source license used in Section 3, and provide a first tutorial in Section 4. This is followed by a *conceptual* overview in Section 5, which itself is sandwiched by two examples of increasing complexity (Sections 4 and 6). With the big picture laid out, Blang's declarative syntax and structure are formalized and detailed in Section 7. A *cheatsheet* highlighting the key ideas discussed in the preceding sections is summarized in Section 8. The sections have been arranged in what the authors believe to be a pedagogical format. However, readers may find it helpful to first skim the sections consisting of examples (Sections 4, 6, and 8). The remaining sections are more advanced, but are nonetheless helpful for drawing context and understanding the motivation behind Blang's design. Section 9 illustrates and discusses a key feature of Blang: the creation of custom data structures and custom samplers. Section 10 introduces Blang's software development kit (SDK), which can be used to implement complex models and assist in testing the correctness of implementations. Section 11 consists of design patterns. Finally, Section 12 describes Blang's architecture and inference algorithms as a whole.

2. Goals

Blang's purpose is to provide Monte Carlo approximations of posterior distributions arising in Bayesian inference problems. The design of the language and its software development kit is guided by the following high-level goals:

Correctness: Bayesian inference software is notoriously difficult to implement. An example from the tip of the iceberg is shown in [Geweke \(2004\)](#), which identifies software bugs and erroneous results in earlier published studies. We address this issue using a marriage of statistical theory and software engineering methodology, such as compositionality and unit testing.

Ease of use: Blang uses a familiar BUGS-like syntax and it is designed to be integrated well in modern data science workflows (input in tidy format, [Wickham 2014](#), samples output in tidy format).

Generality: As a programming language, Blang is Turing-complete and equipped with an open type system, as well as facilities to quickly develop and test sampling algorithms for new types. By an open type system, we mean that the set of types is not limited to integers and real numbers, and can be arbitrary classes. Blang does not fully automate the process of posterior sampling from user-defined types but instead greatly facilitates the development, composition and sharing of custom sampling algorithms.

Computational scalability: The language is designed to ensure that state-of-the-art Monte Carlo methods can be utilized. In particular, we made certain trade-offs to ensure that a well-behaved continuum of distributions can be automatically created. This is complemented with methods that extend existing PPL strategies to the combinatorial space, for example a code scoping analysis to discover sparsity patterns with arbitrary types, as well as built-in support for parallelization to arbitrary numbers of cores.

3. License, source, version and documentation availability

Blang is free and open source. The language and SDK are available under a permissive BSD 2-Clause license. The relevant GitHub repositories are linked at <https://github.com/UBC-Stat-ML/blangDoc>. Online documentation is available at <https://www.stat.ubc.ca/~bouchard/blang/>, including Javadoc pages at <https://www.stat.ubc.ca/~bouchard/blang/Javadoc.html>.

4. Tutorial

This section aims to introduce readers to Blang by presenting a minimal working example. We begin with instructions for performing inference on a simple model using the command-line interface (CLI). Realistic applications are demonstrated in Sections 6 and 9. Advanced tutorials can be found in Appendix A.

4.1. Installing Blang’s command-line interface

We provide instructions here for installing and using Blang via the CLI. Alternative Blang interfaces include an integrated development environment (IDE) as well as a Web interface, both detailed in Section 10.1. Instructions are also available from the documentation website (<https://www.stat.ubc.ca/~bouchard/blang/>) under the link Tools. Additionally, an R ([R Core Team 2022](#)) and Python ([van Rossum et al. 2011](#)) interface to Blang are currently under development.¹

¹The interfaces and associated instructions will be hosted on <https://github.com/UBC-Stat-ML>.

The prerequisites for the CLI installation process are:

1. A UNIX-compatible environment running `bash` or `zsh`. This includes, in particular, Mac OS X, Linux, and Windows Subsystem for Linux.
2. The `git` command.
3. The Java Software Development Kit (SDK), version 11, 13, or 15. Other versions of Java may be incompatible with the version of Xtext our software builds upon.² The Java runtime environment is necessary, and the *runtime environment* is not sufficient, as compilation of models requires compilation into the Java Virtual Machine. Type `javac -version` to test if the Java SDK is installed. If not, the Java SDK is freely available at <https://openjdk.java.net/>.
4. Optionally, if automatic plotting of posterior distributions, trace plots, diagnostics, etc. is required, R as well as the packages `dplyr` (Wickham, François, Henry, and Müller 2021) and `ggplot2` (Wickham 2016) should be installed. In particular, the command `Rscript` should be in the `PATH` variable for the optional plotting functionalities to work correctly.

The following installation process is most thoroughly tested on Mac OS X and Linux, however users have reported installing it successfully on certain Windows configurations (using either Windows Subsystem for Linux or Cygwin).³

To install the CLI tools, input the following commands in a `bash` or `zsh` terminal interpreter:

```
$ git clone https://github.com/UBC-Stat-ML/blangSDK.git
$ cd blangSDK
$ source setup-cli.sh
```

The `git clone` command downloads the `blangSDK` repository, `cd` changes the current working directory, and `source setup-cli.sh` compiles and installs Blang (i.e., updates the `PATH` variable). If the user moves the `blangSDK` folder, the command `source setup-cli.sh` needs to be rerun.

You may now use Blang from any directory by typing `blang` (use lower case for the CLI command as UNIX is case-sensitive).

4.2. Posterior inference

Consider the simplified Doomsday Argument (Carter and McCrea 1983) for modeling the total number of humans that were ever or will ever be born. Denote the estimated number of humans that have been born up to the present time as y , and the total number of humans that were ever or will ever be born (an unknown variable) as z . The Doomsday Argument posits $y | z \sim \text{Uniform}(0, z)$. With a prior belief of what values z can take on encoded as

²Specifically, OpenJDK 8, 11, 13, and 15 have been tested at the time of writing. Java is typically backward compatible, but since the library Xtext performs bytecode manipulations it is more sensitive to versioning than typical Java libraries. Managing and installing several versions of Java is greatly facilitated by the easy to install package `sdkman` available at <https://sdkman.io/>.

³Note also that the Eclipse IDE plug-in does not require a UNIX-compatible environment, see Section 10.1.

```

package toy

model Doomsday {
  param RealVar rate
  random RealVar y
  random RealVar z
  laws {
    z | rate ~ Exponential(rate)
    y | z ~ ContinuousUniform(0.0, z)
  }
}

```

Figure 1: Doomsday model programmed in Blang, `Doomsday.bl`.

an exponential distribution, we update our belief using a PPL to obtain an approximation of the posterior distribution of $z | y$. Using a PPL for such a simple model is excessive but is useful for demonstrating the basic mechanics of Bayesian inference in Blang. Figure 1 shows the entire contents of the Blang file used to code the Doomsday model, `Doomsday.bl`.⁴

The first line is a package declaration, which identifies the package in which the Doomsday model belongs to. The remaining code illustrates four Blang keywords.

- **model**: There should be exactly one `model` per file. The keyword should be followed by an identifier, in this case `Doomsday`. Blang is a case-sensitive language and we use the convention that model names are capitalized.
- **random** and **param** are used to declare model variables. By default, Blang approximates the posterior distribution over the latent `random` variables conditioning on the observed `random` variables.
- Variables need to specify their types. For example, `random RealVar z` is of type `RealVar` and we give it the name `z`. As a convention, types are capitalized and variable names are not.
- Briefly, `random` variables encompass all observed and unobserved random variables. `param` variables encompass all known constants. The distinction is further discussed in Section 5.
- Each model is required to have exactly one `laws` keyword followed by a code chunk surrounded by curly braces, called the *laws block*. The purpose of the laws block is to define joint distributions over the random variables. Here, we show one method to do so, which is inspired by the BUGS notation and its derivatives. For example,

```
y | z ~ ContinuousUniform(0.0, z)
```

denotes the conditional distribution of `y` given `z` is equal to a uniform distribution between 0 and `z`. In contrast to BUGS, we require specification of the random variables that we are conditioning on, here `| z`.⁵

⁴See also https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/Doomsday.bl. Note that the package statements are different.

⁵There are several motivations behind this design choice deviating from BUGS. Technically, static analysis could identify the list of variables we are conditioning on. However the notation used here is closer to a

From the `project` directory, type the following command, in which we specify that `rate` and `y` are fixed to given values while `z` is unobserved:

```
$ blang --model toy.Doomsday --model.rate 1.0 --model.y 1.2 --model.z NA
```

The same model can be run via the Eclipse IDE (with the `--model` argument omitted), following instructions from Section 10.1,⁶ or via a prepackaged repository of examples:

```
$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example/
$ blang --model jss.Doomsday --model.rate 1.0 --model.y 1.2 --model.z NA
```

```
Compilation {
  ...
} [ ... ]
Preprocess {
  ...
} [ ... ]
Inference {
  ...
} [ ... ]
executionMilliseconds : 1037
outputFolder: ./JSSBlangCode/.../results/all/2019-06-27-14-13-21-RL.exec
```

Samples approximating the posterior distribution of `z` given the observation `y` are outputted in tidy format (Wickham 2014) to `samples/z.csv` located in the directory specified by `outputFolder`.

By default, posterior inference is done in two stages. The first stage, corresponding to the `Initialization` block in the standard output, uses SCM which attempts to automatically identify configurations of positive density. In the second stage, an adaptive non-reversible PT algorithm is initialized from the output of the first stage and performs a series of adaptation rounds, corresponding to the `Round(1/9)` through `Round(9/9)` blocks in the standard output. PT algorithms are known to perform well even in the face of difficult sampling problems such as those arising in multimodal distributions or weakly identifiable models. We describe the inference algorithms and their configuration in detail in Section 12.1.

5. Conceptual overview

We now describe more formally the semantics of our language's core construct: the `model`. The basic notation introduced here will be useful to describe the syntax in full detail in the next section.

mathematical notation used for example in the Bayesian non-parametric literature (e.g., Teh, Jordan, Beal, and Blei 2006; Griffiths and Ghahramani 2011). More importantly however, the explicit conditioning allows us to generalize the notation to handle complex dependencies. This is demonstrated in Sections 11.1 and 11.2.

⁶Instructions hosted on Blang's website will be continually updated, see <https://www.stat.ubc.ca/~bouchard/blang/>.

<pre> model Doomsday { param RealVar rate random RealVar y random RealVar z laws { ... } } </pre>	$\theta = (\text{rate})$ $x = (y, z)$ $\{f_\theta\} = \{\text{Doomsday}(\text{rate})\}$ $\Theta_1 = T_1 = T_2 = \text{RealVar}$
--	---

Figure 2: Blang syntax (left) and corresponding mathematical notation (right).

5.1. Models

A Blang `model` encodes a set of *densities* $\{f_\theta(x) : \theta \in \Theta, x \in T\}$, and hence the distribution of a random object $X : \Omega \rightarrow T$. We use the term *density* in a generalized sense, encompassing discrete, continuous, and mixed models, by allowing it to be defined with respect to customizable reference measures.

We assume $x = (x_1, x_2, \dots, x_n)$ where $n < \infty$ is fixed. Despite n being finite in this formalism, each x_i is permitted to be of random or infinite dimensionality. The *type* or space, in which the x_i 's lie, is denoted by T_i . Hence $x_i \in T_i$ and $x \in T = T_1 \times T_2 \times \dots \times T_n$. We also assume each type T_i is implicitly associated with a default reference measure μ_i . These default choices can be changed using the `is` keyword defined in Section 7.10. Once each reference measure μ_i is given, by definition the densities are turned into distributions as follows:

$$P_\theta(X \in A) = \int_A f_\theta(x) \prod_{i=1}^n \mu_i(dx_i), \quad (1)$$

where A is some event, or more formally, an element of the σ -algebra of T . We also assume a decomposition for the parameters $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ where m is fixed and each coordinate θ_j has its type denoted by Θ_j . Hence, $\theta_j \in \Theta_j$ and $\theta \in \Theta = \Theta_1 \times \Theta_2 \times \dots \times \Theta_m$. We use the terminology *model variables* to refer to x and θ collectively.

To understand how these mathematical concepts translate into Blang syntax, let us relate them via the Doomsday example from Section 4. The correspondence is shown in Figure 2. The variables marked with the `random` keyword are concatenated to form x , while those marked with the `param` keyword are concatenated to form θ .

5.2. Interpretation of laws blocks

The laws block is responsible for computing the point-wise evaluation of $\log(f_\theta(x))$ for any input x and θ . To do so, two methods are supported:

Composite laws use existing Blang models as building blocks to create a new one.

Atomic laws provide an arbitrary algorithm to compute the log density.

Both composite and atomic laws allow the user to express a known factorization of the density

$$f_\theta(x) = \prod_{k=1}^K f^{(k)}(x, \theta). \quad (2)$$

```

model ContinuousUniform {
  random RealVar realization
  param RealVar min
  param RealVar max

  laws {
    logf(min, max) {
      if (max - min <= 0.0) return NEGATIVE_INFINITY
      return - log(max - min)
    }
    logf(realization, min, max) {
      if (min <= realization && realization <= max) return 0.0
      else return NEGATIVE_INFINITY
    }
  }
}

```

Figure 3: Continuous uniform distribution programmed in Blang, `ContinuousUniform.bl`.

Such a factorization can then be used as the basis of automating key aspects of state-of-the-art Monte Carlo methods, such as the construction of a well-behaved continuum of auxiliary distributions and the detection of sparsity patterns. Additionally this factorization enables efficient sampling of latent variables, as only a fraction of factors will require evaluation per variable.

5.3. Interpretation of atomic laws

In the case of an atomic law, for each $k \in \{1, 2, \dots, K\}$, an expression or algorithm is provided to compute the value of factor k in log scale, i.e., $\log(f^{(k)}(x, \theta))$.

For example, consider the continuous uniform distribution, which can be factorized as

$$f_{\theta}^{\text{unif}}(x) = \underbrace{\frac{1}{\theta_2 - \theta_1}}_{f^{(1)}(x)} \underbrace{\mathbf{1}[\theta_1 \leq x \leq \theta_2]}_{f^{(2)}(x)},$$

where $\theta = (\theta_1, \theta_2) = (\text{min}, \text{max})$. Figure 3 shows the `model` defining a `ContinuousUniform` distribution in `ContinuousUniform.bl`⁷ of the Blang SDK.

5.4. Interpretation of composite laws

In the case of a composite law, the decomposition in Equation 2 typically comes from an application of the chain rule. In the Doomsday example, this is just:

$$f_{\theta}^{\text{Dooms}}(x) = \underbrace{\theta_1 \exp(-\theta_1 x_2)}_{\tilde{f}^{(1)}(x, \theta)} \underbrace{\frac{\mathbf{1}[0 \leq x_1 \leq x_2]}{x_2}}_{\tilde{f}^{(2)}(x, \theta)}. \quad (3)$$

To understand composite laws, notice the factors in this decomposition can often be retrieved from another existing `model`. In such a case, we say that a `model`, $\{f_{\theta}^{\text{caller}}(x) : x \in T, \theta \in$

⁷See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/others/ContinuousUniformExample.bl.

Θ }, *calls* another model, $\{f_{\theta'}^{\text{callee}}(x') : x' \in T', \theta' \in \Theta'\}$. This is illustrated in our running example as the `Doomsday` model, the caller, calls the `ContinuousUniform` model, the callee. Consequently this allows us to write the second factor in Equation 3 using the previously defined `ContinuousUniform` model via

$$\tilde{f}^{(2)}(x, \theta) = f_{t(x, \theta)}^{\text{unif}}(s(x)),$$

for $t(x, \theta)$ and $s(x)$ defined as follows.

First, $t : T \times \Theta \rightarrow \Theta'$ is a transformation from the *caller* model's variables into the *callee* model's parameters, in this case $t(x, \theta) = (0, x_2)$. The two entries in the list $(0, x_2)$ correspond to the two `param` variables, `min` and `max`, in the definition of `ContinuousUniform` shown in Section 5.3. We see that the order in which the `param` are declared is important when a `model` is to be used in a composite fashion.

Second, $s : T \rightarrow T'$ is a selection of a subset $i_1, \dots, i_{|x'|}$ of coordinates in x , so that $s(x) = (x_{i_1}, \dots, x_{i_{|x'|}})$. Hence, s selects which of the calling model's random variables are used as the callee model's random variables. Here $s(x) = (x_1)$, where the single entry, (x_1) , corresponds to the `random` variable, `realization`, in the definition of `ContinuousUniform`. Again, if more than one random variable is selected, the order in which they are declared in the callee model determines how they are matched.

Considering now the Blang statement:

```
y | z ~ ContinuousUniform(0.0, z)
```

we see that the left of the pipe symbol, `|`, encodes the selection s , and the expression in parentheses encodes the transformation t .

In summary, the two lines in the laws block of the `Doomsday` model:

```
z | rate ~ Exponential(rate)
y | z ~ ContinuousUniform(0.0, z)
```

have the same interpretation as they would in probability theory. However, our notation can also be extended to useful novel patterns (see Sections 11.1 and 11.2).

5.5. Model tree

Composite laws induce a directed tree over models, where a directed edge denotes a `model` calling another `model`. We call this tree the *model tree*. The root of this tree is called the *root model*.

5.6. Interpretation of generate blocks

In addition to the atomic and composite constructs available to specify a mandatory `laws` block, Blang provides an optional orthogonal way to specify $P_{\theta}(X \in A)$, called a `generate` block. The `generate` block performs *forward simulation*: it takes as input a random seed, $\omega \in \Omega$, and returns $X(\omega)$ such that Equation 1 holds.

The `generate` block is technically redundant, but is crucial to check software correctness by setting up statistical unit tests as described in Section 10.5. It is also used for various

purposes during posterior inference, for example, by providing a form of regeneration in PT, and to initialize SCM samplers.

5.7. Normal form

A laws block containing either only composite laws or only atomic laws is said to be in *normal form*. For example, the laws block in `Doomsday.bl` is in normal form, as it consists of composite and only composite laws. Similarly, the laws block in `ContinuousUniform.bl` is also in normal form, as it consists of atomic and only atomic laws. As a counterexample, the following laws block in a model is not in normal form:

```
z | rate ~ Exponential(rate)
logf(z) {
  return -log(z)
}
```

as it contains both composite and atomic laws. Laws blocks in normal form are useful to automatically construct sequences of annealed distributions, used in certain samplers used by Blang’s runtime architecture (see *Constructing a sequence of measures* in Section 12.1).

A model is said to be in *generative normal form* if it satisfies the following conditions:

1. All models in the model tree are in normal form.
2. All models in the model tree based on atomic laws attached to unobserved variables are equipped with a `generate` block.

Generative normal form is only required if the inference engine is PT or SCM, as samples from the prior are exploited for initialization and/or regeneration.

We show in Section 11.1 how to rewrite a wide range of models into a generative normal form. If a model cannot be written in generative normal form, the user may still apply standard MCMC methods but not the more advanced PT and SCM schemes.

5.8. From Blang models to posterior inference

Any Blang model can be transformed into a posterior inference computer program. The inputs of this computer program consist of variables in the root model. All `param` variables in the root model become required inputs. In contrast, `random` variables in the root model can either be specified or left missing as latent. The target posterior distribution is then defined as the distribution of latent random variables given the variables that have been given an input value.

6. Tutorial: A complete example

We illustrate an example of posterior inference for a Gaussian mixture model (GMM). We highlight and briefly discuss key components in implementing a model, and showcase a series of post-processed statistics and plots. After a formal introduction of the syntax (Section 7), we will return to this example in the form of a summary in Section 8.

```

package jss.gmm

model MixtureModel {

  random List<RealVar> y
  param Integer n ? : y.size
  param Matrix a ? : fixedVector(1.0, 1.0)
  random List<IntVar> z ? : latentIntList(n)

  param Integer K ? : 2
  random Simplex pi ? : latentSimplex(K)
  random List<RealVar> mu ? : latentRealList(K)
  random List<RealVar> sd ? : latentRealList(K)

  laws {

    pi | a ~ Dirichlet(a)

    for (int k : 0 ..< K) {
      mu.get(k) ~ Normal(0.0, 100.0)
      sd.get(k) ~ ContinuousUniform(0.0, 10.0)
    }

    for (int i : 0 ..< n) {
      z.get(i) | pi ~ Categorical(pi)
      y.get(i) | mu, sd, IntVar k = z.get(i)
        ~ Normal(mu.get(k), pow(sd.get(k), 2.0))
    }
  }
}

```

Figure 4: Finite Gaussian mixture model programmed in Blang, `MixtureModel.bl`.

Consider the following model:

concentration	$\alpha = [1, 1]$
proportions	$\pi \mid \alpha \sim \text{Dirichlet}(\alpha)$
labels	$z_i \mid \pi \sim \text{Categorical}(\pi)$
means	$\mu_k \sim \text{Normal}(0, 10^2)$
standard deviations	$\sigma_k \sim \text{Uniform}(0, 10)$
observations	$y_i \mid \mu, \sigma, z_i \sim \text{Normal}(\mu_{z_i}, \sigma_{z_i}^2)$

for $i \in \{1, 2, \dots, n\}$ and $k \in \{1, 2\}$.

Figure 4 shows the Blang file `MixtureModel.bl` that encodes a GMM.⁸ We begin by declaring variables as we did in the Doomsday model. In addition to declarations, we initialize them to their respective latent types. Default initializations are expressed using `?:` followed by an expression in a syntax called XExpression described in detail in Section 7.11.⁹ Default initializations can be overridden from the CLI (command line interface). We discuss this

⁸See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/gmm/MixtureModel.bl. Complete and commented implementations in this section are available in the reproduction materials located in the directory `reproduction_materials/example`.

⁹Those familiar with Java can think of XExpressions as “shorthand Java” for now.

mechanism in detail in Section 7.7. In this example, interpret initializations as creating instances of latent objects.¹⁰ A list of data types available for latent variables can be found in Figures 10 and 11.

In the next code block, the `laws` block, we declare the distribution of each latent variable. We use `for` loops to encode a set of declarations. For example, the following two implementations are equivalent:

```
for (int k : 0 ..< 2) {
  mu.get(k) ~ Normal(0.0, 100.0)
  sd.get(k) ~ ContinuousUniform(0.0, 10.0)
}
```

and

```
mu.get(0) ~ Normal(0.0, 100.0)
mu.get(1) ~ Normal(0.0, 100.0)
sd.get(0) ~ ContinuousUniform(0.0, 10.0)
sd.get(1) ~ ContinuousUniform(0.0, 10.0)
```

To perform posterior inference on `MixtureModel` based on observed y_i 's, we invoke the following commands in the CLI:

```
$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example
$ blang --model jss.gmm.MixtureModel \
> --model.y file data/obs1.txt \
> --engine PT \
> --engine.nChains 36 \
> --engine.nScans 30000 \
> --postProcessor DefaultPostProcessor
```

```
Preprocess {
  ...
} [ ... ]
Inference {
  ...
} [ ... ]
Postprocess {
  Post-processing allLogDensities
  Post-processing energy
  Post-processing z
  Post-processing logDensity
  Post-processing mu
  Post-processing nOutOfSupport
  Post-processing pi
```

¹⁰In Section 9, we create a constructor for objects of type permutation. Its application is helpful in painting a bigger picture on how these latent objects are used behind the scenes.

Argument	Description
<code>--model.y file</code>	Specifies the file path to a newline-separated file with y 's values.
<code>--engine PT</code>	Specifies parallel tempering as the inference algorithm.
<code>--engine.nChains</code>	Controls the number of (annealed) parallel Markov chains.
<code>--engine.nScans</code>	Controls the number of posterior samples to draw.
<code>--postProcessor</code>	Specifies the post-processor.

Table 1: Summary of the key CLI arguments for the example.

```

Post-processing sd
MC diagnostics
} [ ... ]
executionMilliseconds : ...
outputFolder ../JSSBlangCode/.../all/2020-12-31-23-59-03-N0PvDjdc.exec

```

In this example, `obs1.txt` is a newline-separated file formatted as follows:

```

3.2
-0.3
1.7
:

```

More generally, information on the format used to input data can be obtained by *appending* `--help` to the command line arguments (the command line help is contextual, so the information given by appending `--help` to the model and inference engine specific arguments will be more detailed than only using `blang --help`). A more sophisticated method to input data, based on the *plate notation*, is discussed in Section 10.4. We briefly summarize the key CLI arguments for the example in Table 1.

The details of how the `--engine` arguments influence the performance of inference are discussed in Section 12.

All experiment outputs are stored in a `results` directory, within the working directory in which the Blang CLI command is called. Generally, there are three categories of outputs: samples (raw output), post-processed statistics/plots (summaries of the samples), and monitoring statistics/plots (to assess the quality of the posterior approximation). Options for post-processing are handled via the `--postProcessor` runtime argument, accepting `DefaultPostProcessor` or `NoPostProcessor` as arguments. Again use `--postProcessor DefaultPostProcessor --help` for more information.

Currently, the `DefaultPostProcessor` option produces trace and density plots,¹¹ and provides summary statistics including highest density credible intervals (HDIs, constructed using the method described in [Chen and Shao 1999](#)) and effective sample size (ESS) estimates (based on a numerically robust version of the \sqrt{n} -size batch estimator described in [Flegal and Jones 2010](#)). Type information is used to select appropriate plotting strategies (e.g., probability mass functions for `IntVar` types, density estimates for `RealVar`). Examples of summary statistics for `MixtureModel`'s parameters are shown in Table 2, and can be found under the directory summaries in `results/latest`.¹² Notice the posterior summaries are

¹¹The `DefaultPostProcessor` requires R as well as the packages `dplyr` and `ggplot2`.

¹²Numerical values are truncated to fit in the page width.

index	parameter	mean	sd	min	median	max	HDI.lower	HDI.upper
0	mean	0.53	1.75	-19.87	0.96	23.03	-1.38	2.30
1	mean	0.57	1.76	-21.58	1.37	22.48	-1.36	2.34
0	sd	0.92	0.58	0.02	0.85	10.00	0.35	1.34
1	sd	0.92	0.62	0.06	0.85	9.82	0.35	1.32
0	pi	0.50	0.24	0.00	0.49	1.00	0.18	0.82
1	pi	0.50	0.24	0.00	0.51	1.00	0.18	0.82

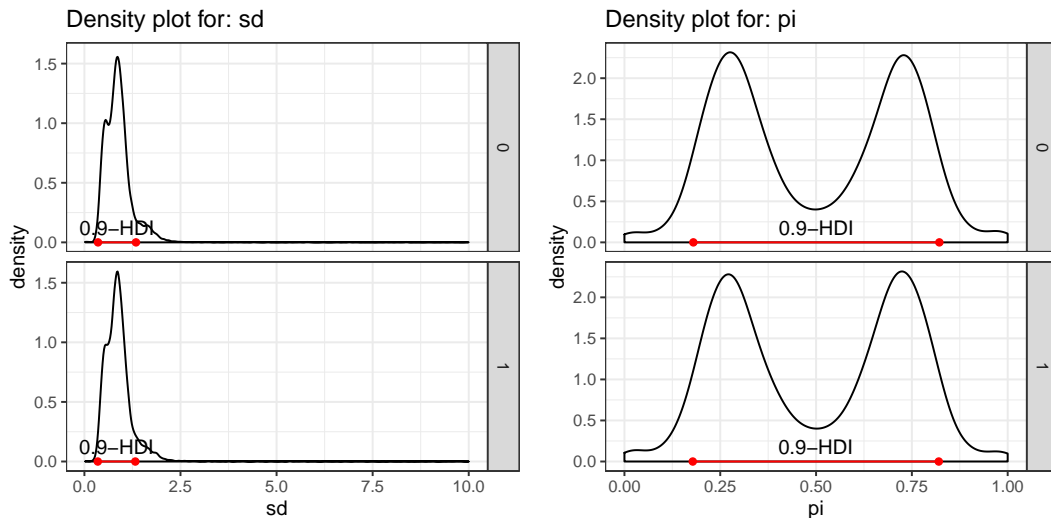
Table 2: Summary statistics for the `MixtureModel`'s parameters.

Figure 5: Posterior density plots for a subset of random variables in the GMM. The facets (rows) are indexed by the mixture components. Left: standard deviation parameters. Right: mixture proportion parameters. The two pairs of nearly-identical plots are indicative of successful label switching, showing that the multimodal posterior distribution is well approximated. By default, the 90% highest density interval is underlined in red.

nearly identical for the two mixture components. Similarly, the marginal posterior plots in Figure 5 also exhibit this symmetry. This symmetry is to be expected in this example: it arises from the unidentifiability of the GMM parameters known as label switching (Jasra, Holmes, and Stephens 2005). Here the inference engine used, an adaptive non-reversible parallel tempering algorithm (abbreviated PT), is capable of capturing this symmetry despite the high-dimensional multimodality involved (the z_i 's of all variables have to be flipped to switch modes).

Another statistic that is often of interest is the normalization constant (also known as model evidence, or marginal likelihood). The logarithm of this value is automatically output in `logNormalizationEstimate.csv`. The various methodologies available to estimate the log normalization constant are discussed in Section 12.4. Figure 7 illustrates the progression of estimates across PT adaptation rounds.

Output files for diagnosing and monitoring the performance of inference algorithms are also produced. We will describe them in Section 12.1.

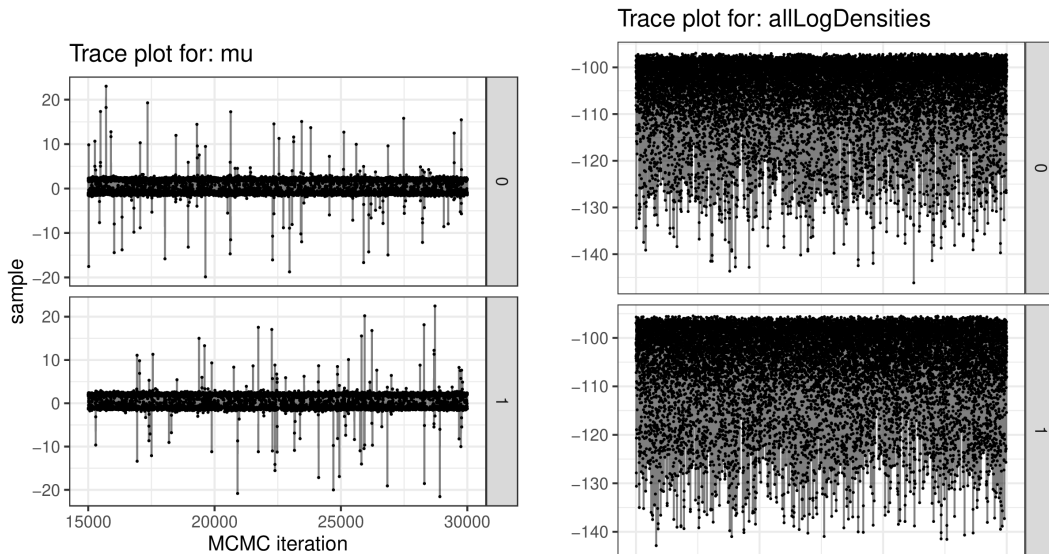


Figure 6: Left: Trace plot for cluster-specific location parameters. The two clusters are shown as facets. Right: log densities for two of the 36 tempered chains used in PT. The x -axis has been clipped in the plot on the right but it is identical to the x -axis of the plot on the left. Notice that the “jumps” between modes are densely distributed along the traces, i.e., they occur very frequently in this example. Other diagnostics produced will be discussed in Section 12.1.

7. A complete tour of Blang’s syntax

In this section we provide a more systematic survey of the Blang language. The formal definition of the language can be accessed in the `blangDSL` repository at <https://github.com/UBC-Stat-ML/blangDSL>.

7.1. Project organization

Blang projects are composed of three types of files: Blang files (`.bl`), Xtend (Xtend 2019) files (`.xtend`), and Java files (`.java`). This section is devoted to the syntax of Blang files. Xtend and Java files are used to create supporting code for non-standard data types, samplers, and user-defined functions. The user can choose either Xtend or Java for creating supporting code. For users not familiar with Java, we recommend using Xtend because its syntax is consistent with Blang’s syntax. This is a consequence of both languages being constructed with the Xtext language development framework.

7.2. Interoperability with Java

Blang, Xtend and Java are seamlessly interoperable as the first two are transpiled into Java. More precisely, any Java type can be imported and used in Blang, and any model defined in Blang can be imported and used in Java with no extra work needed.

As such, types in Blang are equivalent to *Java types*, a terminology that encompasses Java classes, interfaces, primitives, enumerations and annotation interfaces. At a high level, a type

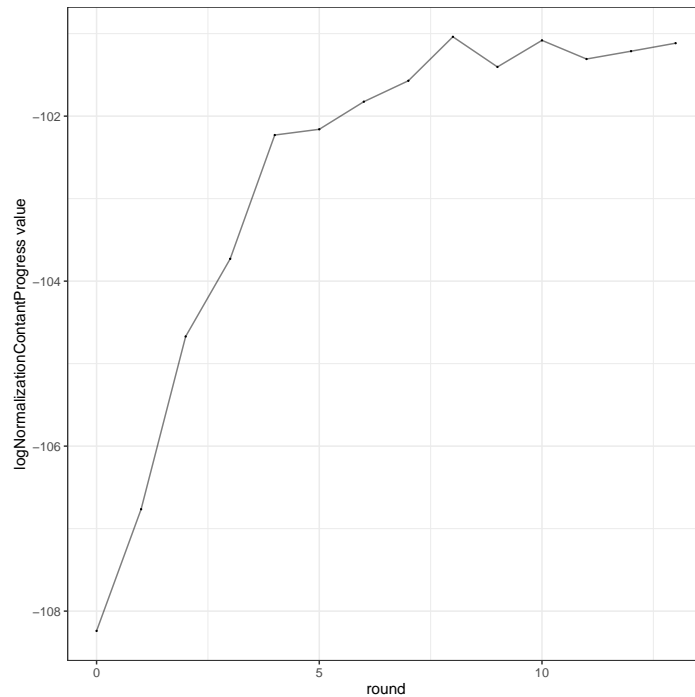


Figure 7: Log normalization constant estimates across adaptation rounds when the PT algorithm is used. The fact that these estimates plateaued supports that the allocated computational budget is sufficient for this inference task.

can be thought of as a group of *objects* (chunks of computer memory) that satisfy a certain set of properties (for example, they all support being passed in a certain function). We do not assume prior knowledge of the Java language, in fact, Blang and Xtend syntax is often simpler compared to Java's.

7.3. Comments

Single line comments use the syntax

```
// some comment spanning the rest of the line.
```

Multi-line comments use

```
/*
many commented lines
can go here
*/
```

7.4. Blang models: High-level syntax

A Blang file is organized as shown in Figure 8. We briefly describe each code block as follows: package statements are responsible for defining the package in which a Blang model belongs to. Import statements are responsible for importing classes, functions, and models from

```
// package and import statements

model NameOfMyModel {

    // variables declarations

    laws {
        // laws declaration
    }

    generate(nameOfMyRandomObject) {
        // generate block
    }
}
```

Figure 8: Code illustrating the structure of a Blang file, `NameOfMyModel.bl`.

other packages. The variables declarations block is responsible for declaring model variables, i.e., observed (constant) variables, latent variables, unknown parameters, known (constant) parameters. The laws block is used to declare the probability distribution associated with each of the (random) model variables (see Section 7.8). The optional generate block is used for forward sampling from the model (see Section 7.9). It will also be helpful to keep in mind that XExpressions (to be introduced) are imperative, while laws blocks are declarative. Declarative code blocks do not have a notion of order, in other words, permuting the order of two statements will have no observable effect on the program.

In the remainder, if a string such as `NameOfMyModel` contains the substring “My”, or has an integer as suffix, it refers to an identifier that should be tailored to the context of the model being written.

Blang is case-sensitive. Identifiers (model names, variable names, etc.) should start with a letter and only use letters, numbers, and underscores. Furthermore, as a convention we encourage users to capitalize model names.

7.5. Packages and imports

The *packages* construct deals with the rare, but unavoidable, situation of wanting to use code from two developers that used the same name for a Blang model. Package declarations will disambiguate the two.

Packages in Blang work the same as in Java, and precede `import` statements. To declare a Blang model as part of a hierarchical group of related code, place the following declaration at the very beginning of the Blang file:

```
package myOrganization.myPackageName
```

This *package declaration* line is optional but recommended if you plan to share your code. The dot in `myOrganization.myPackageName` denotes a hierarchical organization going from broader to more specific from left to right. As a convention, package names are generally not capitalized.

To use another Blang model called `AnotherModel` from a package named `some.other.pack`, we can use `import` statements of the form:

```
import some.other.pack.AnotherModel
```

after the package declaration line. The same syntax can be used to import Java or Xtend classes, where `import static` is used to import a function, while a standalone `import` statement is used for types.

Package declarations effectively enable users to refer to specific objects of a package explicitly through import statements. In the example below we see why this would be useful. Suppose our model requires two data types from `package1` and `package2`, each of which contain an identically named but different implementation of `DupedType`. In the unlikely event of having to use two types with duplicated names within the same file, importing should be avoided (i.e., do neither `import package1` nor `import package2`). Instead each instance of the type should be prefixed with the package name within the code. For example, consider the following code in `MyModel.bl`:

```
model MyModel{
  random package1.DupedType var1
  random package2.DupedType var2
  ...
}
```

In contrast, here is an example of what *not* to do:

```
import package1.DupedType
import package2.DupedType

model MyModel{
  random DupedType var1
  random DupedType var2
  ...
}
```

A related construct is the extension import mechanism, described in more detail in Section 7.11.

7.6. Automatic imports in Blang files

Any Blang file automatically imports:¹³

- all the types in the following packages:
 - `blang.core`,
 - `blang.distributions`,
 - `blang.io`,
 - `blang.types`,
 - `blang.mcmc`,
 - `java.util`,
 - `xlinear`
- all the static functions in the following files:
 - `xlinear.MatrixOperations`,

¹³The relevant Javadocs can be found at <https://www.stat.ubc.ca/~bouchard/blang/Javadoc.html>.

```

bayonet.math.SpecialFunctions,
org.apache.commons.math3.util.CombinatoricsUtils,
blang.types.StaticUtils

```

- as static extensions all the static functions in the following files:

```

xlinear.MatrixExtensions,
blang.types.ExtensionUtils,
blang.distributions.Generators

```

7.7. Model variables

Model variables encompass all observed (fixed) variables, latent variables, unknown parameters, and known (constant) parameters in a statistical model. Model variables are declared using one of two methods, declared with no default initialization:

```

random Type1 name1
param Type2 name2

```

or with default initialization:

```

random Type3 name3 ?: XExpression1
param Type4 name4 ?: XExpression2

```

Observed and latent random variables are declared with `random`, while parameters are declared with `param` (see Section 5.1). The initialization blocks, denoted by `XExpression1` and `XExpression2`, are imperative blocks of code used to provide default values in the absence of CLI arguments. For example:

```

random Double abc ?: {
    val x = 123.0
    return exp(x)
}

```

The expressions in initialization blocks are constructed with so called *XExpressions*. XExpressions are introduced in more detail in Section 7.11 and are used to construct several aspects of Blang programs. For now, think about XExpressions as chunks of code (lists of statements or expressions) capable of performing arbitrary computations (loops, conditionals, creating temporary variables, calling other functions, etc.), and returning one value. The statements or expressions in a block can be terminated by a newline or by a semicolon.

If the block contains only one expression, the brackets can be omitted:

```

random Double abc ?: exp(123.0)

```

Initialization blocks can use values of previously listed variables. If a CLI argument is provided, then the initialization block will be overridden by it.

7.8. Laws block

Laws blocks are used to *declare* the (conditional) probability distribution associated with each random variable. Note that unlike common programming languages used today for data

analyses such as Python and R, the `laws` block is declarative. In particular, the interpretation of a model is invariant to the order in which the individual laws are declared in the code.

Composite laws

Described conceptually in Section 5.4, composite laws have the following syntax in Blang:

```
variableExpression1, variableExpression2, ...
| conditioning1, conditioning2, ...
~ MyDistributionName(argumentExpression1, argumentExpression2, ...)
```

For example:

```
y | mu, variance ~ Normal(mu + 123, variance)
```

where `variableExpression1`, `conditioning1`, `conditioning2`, `argumentExpression1`, and `argumentExpression2` correspond to `y`, `mu`, `variance`, `mu + 123`, and `variance` respectively. `MyDistributionName` refers to another Blang model. Each element in `argumentExpression1`, `argumentExpression2`, ... is matched from left to right in the same order as the `param` variables are declared in the model `MyDistributionName`.

The list (`argumentExpression1`, `argumentExpression2`, ...) corresponds to the transformation $t : T \times \Theta \rightarrow \Theta'$ in the notation used in Section 5.4. This is implemented by allowing each element in `argumentExpression1`, `argumentExpression2`, ... to be an `XExpression` which is recomputed each time the value of the density $f_{\theta}(x)$ is queried; the expressions `argumentExpression1`, `argumentExpression2`, ... are compiled to lambda expressions. Continuing with the example above, `mu + 123` will be computed at every iteration of an MCMC algorithm (when factors dependent on μ are required). In other probabilistic programming languages, these expressions are often referred to as deterministic nodes/variables (in `RevBayes` and `BUGS` for example).¹⁴

Each element in `variableExpression1`, `variableExpression2`, ... is matched from left to right in the same order as the `random` variables are declared in model `MyDistributionName`. To relate this to Section 5.4, the list `variableExpression1`, `variableExpression2`, ... corresponds to the output of the selection function $s : T \rightarrow T'$. This is implemented by allowing each `variableExpression` to be an `XExpression` which is executed only once, at initialization time. Often this `XExpression` is only a variable name, but it could also be an expression selecting an entry in a list or vector.

The conditioning block, `conditioning1`, `conditioning2`, ... is used to restrict what can be accessed by the transformation t . This is called the *scope* of the transformation t . It is useful to restrict the scope as much as possible since this restriction induces sparsity patterns in the model. Sparsity is then exploited by our efficient inference algorithms.

Specification of the scope is implemented as follows. Each item within `conditioning1`, `conditioning2`, ... can take one of two possible forms. First, it can be one of the variable names declared via the keyword `random` or `param`. For example, this first method is used in all conditionings of the Doomsday model (see Figure 2).

The second method to specify a conditioning is as follows:

¹⁴In contrast to these other languages, these deterministic nodes cannot be straightforwardly named and traced at the moment. We are investigating ways to incorporate this feature in future releases.

```
variableExpression1, variableExpression2, ...
| MyType myConditioningVariable = XExpression1, ...
  MyDistributionName(argumentExpression1, argumentExpression2, ...)
```

where `MyType` is a type, `myConditioningVariable` is a local variable that exists only for the declaration of `variableExpression1`, `variableExpression2`, ...'s law. The code in `XExpression1` has access to all model variables. For example:

```
y | RealVar mu = manyMus.get(0) ~ Normal(mu, 1)
```

The `XExpression1` code is executed only once at initialization. We show in Section 11.2 an example of the typical use case for this initialization process, where in a model for a Markov chain, this initialization is simply to select, in a list of random variables, the variable corresponding to the previous time step.

Atomic laws

Informally, atomic laws are used to compute factors, and are the building blocks for composite laws. Described conceptually in Section 5.3, atomic laws have the following syntax in Blang:

```
laws {
  logf(expression1, expression2, ...) { XExpression }
}
```

For example, $x \sim \text{Normal}(\mu, \sigma^2)$ (realization $\sim \text{Normal}(\text{mean}, \text{variance})$) would have the following encoding in `Normal.bl`:

```
laws {
  logf(mean, variance, realization) {
    if (variance < 0.0) return NEGATIVE_INFINITY
    return (- 0.5 * log(2*PI)
            - 0.5 * log(variance)
            - 0.5 * pow(mean - realization, 2) / variance)
  }
}
```

It is recommended to separate factors with as few arguments together as possible, as this will help the runtime architecture determine dependencies and avoid redundant computation. For example, the `Normal.bl` implementation is recommended to be factorized as:¹⁵

```
laws {
  logf() {
    - 0.5 * log(2*PI)
  }
  logf(variance) {
    if (variance < 0.0) return NEGATIVE_INFINITY
  }
}
```

¹⁵See <https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/distributions/Normal.bl>.


```

    return - 0.5 * log(variance)
  }
  logf(mean, variance, realization) {
    if (variance < 0.0) return NEGATIVE_INFINITY
    return - 0.5 * pow(mean - realization, 2) / variance
  }
}

```

Recall that in Section 5.3, each atomic law was denoted as $\log(f^{(k)}(x, \theta))$. Here the list `expression1, expression2, ...` is used to restrict the scope of $f^{(k)}$, with the same motivation and mechanism as for composite laws, described in the last section. Each item in the list `expression1, expression2, ...` follows the same syntax as the items in `conditioning1, conditioning2, ...` also described in the last section.

The `XExpression` is responsible for computing the numerical value of $\log(f^{(k)}(x, \theta))$, and as such, should return a value of type `Double`. The `XExpression` is recomputed each time the value of the density $f_{\theta}(x)$ is queried.

Declarative loops

In practice, the factorization in Equation 2 may have a large number of factors. To assist the user in declaring these factors, we provide a “declarative loop” construct:

```
for (MyIteratorType myIteratorName : XExpression) { ... }
```

This will repeat all the declarations inside `...` be they atomic or composite. Loops can be nested with the expected cross product behavior.

The `XExpression` should return an object of type `java.lang.Iterable`. Some important loop idioms:

- Simple loop from 0 (inclusively) to 10 (exclusively):

```
for (Integer i : 0 ..< 10) { ... }
```
- Loops based on a `Collection`, which offer a wide choice of data structures via Java’s SDK or Google’s Guava project.¹⁶
 An example iterating over a power set:¹⁷

```
for (Set<Integer> s : (0 ..< 5).powerSet) { ... }
```
- Loops based on Xtend’s or Java’s utilities.¹⁸ An example iterating over the first four even integers:

```
for (Integer i : (0 ..< 10).filter[it % 2 == 0]) { ... }
```

 The keyword `it` is explained in Section 7.11.

¹⁶See <https://docs.oracle.com/javase/tutorial/collections/> and <https://github.com/google/guava/wiki/CollectionUtilitiesExplained>.

¹⁷This requires the import line `import static extension com.google.common.collect.Sets.powerSet`.

¹⁸Documentation for Xtend’s utilities is available at https://www.eclipse.org/xtend/documentation/203_xtend_expressions.html and documentation for Java’s streams is available at <https://docs.oracle.com/javase/tutorial/collections/streams/>.

- The `Plate` data structure supplied by the Blang SDK is described in more detail in Section 10.4.

The current runtime infrastructure assumes that the `XExpression` specifying the range should not be random, in particular, it should not change during sampling. As such it is only computed at initialization. Therefore, declarative loops, which *surround* atomic and composite laws, are different than the loops *within* `XExpressions`. Although Blang does not currently have built-in data types for sampling of infinite dimensional objects, they can be handled by creating dedicated types and/or using `XExpression` loops *inside* a `logf` block.

7.9. Generate block

The generate block is responsible for the forward generating mechanism of a model. This is optional in that it is only required when more sophisticated inference algorithms are desired, as discussed in Section 5.6. An important distinction from `laws` blocks is that `generate` blocks are imperative. Furthermore, they are not referentially transparent as random variables will be modified in-place. We formalize the syntax used to encode the generate block introduced conceptually in Section 5.6:

```
generate(myRandomSeed) {
  XExpression
}
```

The argument `myRandomSeed` is the name of an input object of type `java.util.Random` (the type declaration for this input is skipped since this is the only possible type allowed). To connect this syntax with its interpretation described in Section 5.6, the input argument can be thought as an outcome $\omega \in \Omega$, from which the `XExpression` should form the realization $X(\omega)$.

If the `model` has exactly one `random` variable of type `IntVar` or `RealVar`, then the `generate` block should return an `int` or `double` respectively, corresponding to the new realization. Otherwise, the generate block should modify the `random` variable(s) in-place. The special case for univariate `IntVar` and `RealVar` is just syntactic sugar: under the hood, generated code uses the returned realization to modify the single variable to be sampled in-place.

7.10. Latent random variables and their reference measures

Each type of `random` variable which we would like to be latent is required to declare one or more sampling algorithms. This is achieved by adding the following *type annotation* in the Xtend or Java class for that data type:

```
@Samplers(MySampler1, MySampler2, ...)
class MyDataType {
  ...
}
```

Here each item in the list `MySampler1`, `MySampler2`, ... should be subtypes of the interface `Sampler`.¹⁹

¹⁹<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/mcmc/Sampler.html>

```

...
model Dirichlet {
  random Simplex realization
  param Matrix concentrations

  laws {
    logf(concentrations, realization) {
      ...
    }
    realization is Constrained
  }

  generate(rand) {
    ...
  }
}

```

Figure 9: Dirichlet distribution programmed in Blang, `Dirichlet.bl`.

Implicitly, the samplers associate a default reference measure to the latent `random` variables. In some models, it may be necessary to overwrite these default reference measures for a particular `random` variable. In such cases, Blang provides a mechanism to change them by adding in the `laws` block, a line of the following form:

```

laws {
  ...
  myVariableName is Constrained
}

```

In the above, `myVariableName` refers to the `random` variable name for which the default reference measure is to be changed, and `Constrained` can also be any class which implements `Factor`.²⁰ Effectively, the intended behavior is to disable samplers which would be inoperative with the alternate choice of reference measure.

To illustrate this necessity, consider a K -dimensional Dirichlet distributed random variable (i.e., $p = (p_1, p_2, \dots, p_K)$). By default, Blang would automatically designate slice samplers for each of the coordinates p_1, p_2, \dots, p_K , as they are of type `RealVar` variables. However, because of the simplex constraint requiring $\sum_{i=1}^K p_i = 1$, this would lead to proposal rejections almost surely. The keyword `Constrained` is used to prevent this automatic assignment of ineffective or incorrect samplers.²¹ Thus for the Dirichlet distribution, we require the line `realization is Constrained` to disable each coordinate's default sampler, as seen in Figure 9.

Having disabled default samplers, the next logical step is to ensure the variable (of type `Simplex`) has an appropriate sampler of its own. The reader is recommended to return to this section after building familiarity with Sections 9 and 12.7, where the details of creating custom samplers are discussed. A very high-level introduction to creating samplers is discussed here only to highlight the role of the constraint mechanism (if and when required) within samplers.

²⁰<https://www.stat.ubc.ca/~bouchard/blang/javadoc-dsl/blang/core/Constrained.html>, <https://www.stat.ubc.ca/~bouchard/blang/javadoc-dsl/blang/core/Factor.html>

²¹Technically this could be achieved by creating a `simplex` type that is constructed without referencing a `RealVar` type, but this is cumbersome as it is natural to use standard matrix objects within a `simplex` implementation.

```

...
class SimplexSampler implements Sampler {
    @SampledVariable DenseSimplex simplex
    @ConnectedFactor List<LogScaleFactor> numericFactors
    @ConnectedFactor Constrained constrained

    override void execute(Random rand) { ... }
    ...
}

```

Figure 10: Snippets of the simplex sampler programmed in Xtend, `SimplexSampler.xtend`.

Consider Blang’s implementation of a simplex sampler shown in Figure 10. The annotation `@SampledVariable` informs Blang that the `simplex` field is the variable to be sampled in-place. Here we focus on the line `@ConnectedFactor Constrained constrained` which signals that it is appropriate to use this sampler, even in the presence of a constrained factor connected to the sampled variable in the factor graph. In contrast, the default sampler for real variables, `RealSliceSampler`, does not have `@ConnectedFactor Constrained constrained` stating that it is not able to accommodate sampling of the variable when it is connected to such factor. Again refer to Section 12.7 for a more detailed discussion.

The function `execute` samples `simplex` in-place, or in other words, mutates the variable as an update. The implementation details of `execute` are not important for the discussion of the constraint mechanism, and are thus hidden.

In short, to change a reference measure for a variable, a user should first disable the sampler for a variable by declaring `myVariableName is Constrained` in the `laws` block of a Blang model. Then create a sampler, and annotate a field of type `Constrained` with `@ConnectedFactor` to signal that it can handle this type of constraint.

A more refined typology of constraints can be built by the user, simply by creating subtypes of `blang.core.Factor`. Those used in the standard library might also be refined in future releases.

7.11. XExpressions

Syntax for XExpressions is provided by the Xtext language engineering framework. XExpressions are imperative expressions. Thus the `logf`, `generate`, and variable initialization blocks for example are imperative, while `laws` blocks are declarative.

Here we highlight key aspects commonly used in Blang programs. We refer the reader to the Xtext documentation for more information.²²

XExpressions can be either a single instruction as in the *argument* of the following Exponential composite law:

```
y | a, b, x ~ Exponential(exp(a * x + b))
```

or there can be several instructions nested in braces, with the last one providing the return value, as in this equivalent version of the above code

²²The documentation page can be found at <https://www.eclipse.org/Xtext/documentation/>.

```
y | a, b, x ~ Exponential({
  val product = a * x
  exp(product + b)
})
```

Types

We classify types into three main categories: primitives, object references, and array references. The most common primitives are `boolean`, `int`, and `double`.²³ Object references can be thought of as an annotated address to a memory location, possibly `null`. Lastly, array references are rarely used directly in `Blang`. Instead, arrays are typically encapsulated in more convenient data structures.

Literals

Examples of expressions that create constants of type ...

- `boolean`: `true`, `false`
- `int`: `42`, `12000`
- `double`: `1.0`, `1.3e2`, making sure to include the decimal suffix or to use scientific notation.
- String literals: either via `"A"`, or `"This version allows "quotes inside" and more"`.
- type literals: `MyType`, which is equivalent to Java's `MyType.class`.
- List: `#[true,false]` (note the hash symbol `#` is not a comment as in other languages, it is used to construct lists, sets, and maps).
- Set: `#{"A","C","G","T"}`
- Map: `#{"key1" -> 1, "key2" -> 2}`
- Pair: `"likelihood" -> 1.43` (this example returns type `Pair<String, Double>`; this syntax can be used with arbitrary key and value types).

Declaring variables with XExpressions

Local variables have to be declared at their first occurrence. The main syntax variants to do so are:

```
var int myModifiableInt = 17
var typeInferred = #[1,2,3]
val int myConstantInt = 17
```

²³They have the same characteristics as in `Java`, see <https://docs.oracle.com/javase/tutorial/java/nutsandbolts/datatypes.html> for technical details.

In the example, `var` encodes a variable that is mutable whereas `val` encodes a variable that is immutable. The meaning of immutability is simple to understand in the case of a primitive, but it should be interpreted carefully in the context of references. In the latter, it means that the reference will always point to the same object in the heap, however the internal state of that object might change over time.

In the above, `typeInferred` illustrates that the type can be inferred automatically, in this example a `List<Integer>`.

Conditionals

Conditional expressions have the following form:

```
val String variable = if (condition) value1 else value2
```

Conditional expressions return values depending on a condition, where `condition` evaluates to a boolean. When `(condition)` is `true`, `value1` is returned, otherwise `value2` is returned. The shorthand notation without `else`

```
if (condition) value
```

or

```
if (condition) {
    (value)
}
```

is equivalent to `if (condition) value else null`.

Scope

The *scope* of a variable is defined as the portion of code in which the variable can be accessed. Scoping in Blang is similar to most languages where in order to find the scope of a variable we identify the parent braces and determine the region of the code where the variable can be accessed. For example, a local variable declared within the body of a `for` loop (the regions between curly braces) cannot be accessed outside of the body. If one variable reference is in the scope of several variables declared with the same name, then the innermost braces have priority.

The only exception is the arguments of the atomic and composite laws. Recall our example in Section 7.8 (repeated below),

```
variableExpression1, variableExpression2, ...
| conditioning1, conditioning2, ...
~ MyDistributionName(argumentExpression1, argumentExpression2, ...)
```

These laws require explicit declaration of the variables to include in the scope, where these variables should be identified at the right of the `|` symbol. This design choice is primarily motivated by its flexibility in handling complex dependencies, to be demonstrated in Sections 11.1 and 11.2.

XExpression loops

In addition to allowing loops following the declarative loop syntax, loops within XExpressions allow the number of iterations to be random as well as a few syntactic alternatives:

1. Basic, C-like for loops:

```
for (var IteratorType iteratorName = init; condition; update) {...}
An example of which would be
for (var int i = 0; i <= 10; i++) {...}.
```

2. While loops:

```
while (condition) {...}.
```

Function calls

Functions are called as one would expect: `nameOfFunction(expression1, expression2)` where each element in `expression1` and `expression2` are XExpressions. These expressions are evaluated prior to being passed into the function (i.e., a form of “eager/greedy evaluation”), in order from left to right.

The only exceptions are composite laws, where the evaluation of an argument is delayed at initialization and instead repeated each time the density is evaluated during sampling (i.e., a form of “lazy evaluation”). To see why this is needed, consider a factor declaration of the form $y \mid x \sim \text{Normal}(2 * x, 1)$. Each time this factor is computed during inference, we would like the mean parameter $2 * x$ to be recomputed. One way to think about lazy evaluation in this context is that when the factor graph is created, $2 * x$ is converted into a lambda expression which is computed each time we are computing the value of the normal factor.

In all cases, the actual function call only involves copying a constant size register making these calls very cheap. For primitives, the value of the primitive is copied and therefore the original primitive can never suffer side effects from the call. For object references, the memory address in the reference is copied and hence the original reference cannot be changed, although the object it points to might have its state changed by the function call.

User defined functions

To create supporting functions, the user can create a separate Xtend or Java file. In Xtend, use the following template for the separate file, say in `MyFunctions.xtend` provided below:

```
package my.pack
class MyFunctions {
  def static ReturnType myFunction(ArgumentType1 arg1, ArgumentType2 arg2) {
    // some computation
    return result
  }
}
```

Back to the Blang file being developed, the user can then import the functions into the Blang file using `import static my.pack.MyFunctions.*` allowing us to call `myFunction(arg1, arg2)`.

Extensions

Extension methods provide a kind of lightweight trait, i.e., adding methods to existing classes on demand.

Continuing the same example in the last section, this is done by adding an extension import statement:

```
import static extension my.pack.MyFunctions.myFunction
```

Provided a variable, say `myVar`, of type `ArgumentType1` (the type of the first input argument to the function `myFunction` defined in the previous section), the user can then invoke the function via `myVar.myfunction(arg2)`.

As a concrete example of how this is used to create more readable code, consider a typical `generate` snippet, showing here how a Yule Simon distributed variate can be generated as a mixture

```
generate(rand) {
  val w = rand.exponential(rho)
  return rand.negativeBinomial(1.0, 1.0 - exp(-w))
}
```

This can be equivalently written, more explicitly, as

```
generate(rand) {
  val w = Generators.exponential(rand, rho)
  return Generators.negativeBinomial(rand, 1.0, 1.0 - exp(-w))
}
```

The underpinning of this code is that since Blang automatically imports all functions in `Generators` as extension methods,²⁴ which contain the function:

```
def static double exponential(Random random, double rate)
```

then we can call `rand.exponential(...)` on the variable `rand` of type `java.util.Random`.

Creating objects

An object of type `MyClass` is created by calling `new MyClass(argument1, ...)`. This can be shortened to `new NameOfClass` if there are no arguments. To find which argument(s) are necessary, look for the *constructor* in `MyClass`, which uses the keyword `new` in `Xtend` and the name `MyClass(...)` in Java.

In some libraries, for example in the package we use for linear algebra, `xlinear`, the call to `new` is wrapped inside a static function. In this case, just call the function to instantiate the object. For example, to create a new sparse matrix with 1 000 rows and 10 000 columns, use `sparse(1_000, 10_000)` (automatically imported from `xlinear.MatrixOperations`).²⁵

²⁴<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/distributions/Generators.html>

²⁵<https://www.stat.ubc.ca/~bouchard/blang/javadoc-xlinear/xlinear/MatrixOperations.html>

Using objects

Classes have *instance variables* or *fields*, which are variables associated with objects, as well as *methods*, which are functions associated with the object having access to the object's instance variables. Collectively, fields and methods are called *features*. Features are accessed using the “dot” notation: `myObject.myVariable` and `myObject.myMethod(...)`. When a method has no argument, the call can be shortened to `myObject.myMethod`.

The ability to call a feature is subject to Java visibility constraints. In short, only public features can be called from outside the file declaring a class.

Implicit variable it

The special variable `it` allows users to provide a default object for feature calls:²⁶

```
val it = myObject
doSomething
```

This is merely a shorthand notation for:

```
myObject.doSomething
```

and is used in lambda expressions which we discuss next.

Lambda expressions

A *lambda expression* is a succinct way to write a function without having to give it a name. This construction makes it easy to call functions which take functions as argument (e.g., to apply a function to each item in a list). Since they are so useful, many syntactic shortcuts are available.

The explicit syntax for lambda expressions is:

```
[Type1 argument1, Type2 argument2, ... | functionBody ]
```

For example, to capitalize words in a list, we can use the function `map(myFunction)` which applies `myFunction` to every entry in the list. Here `map()` is the function that takes in another function `myFunction` as an argument. More concretely, we have:

```
#["foo", "bar"].map([String s | s.toUpperCase])
```

When there is a single input argument in the lambda expression (i.e., in the above case `String s`), you can skip declaring the argument, and instead the argument will be assigned to the implicit variable `it` (described in the previous section). This allows us to write:

```
#["one", "two"].map([it.toUpperCase])
```

which further simplifies to:

```
#["one", "two"].map([toUpperCase])
```

²⁶See https://www.eclipse.org/xtend/documentation/203_xtend_expressions.html.

Finally, when the last argument of a function (`map()` in this case) is a function, you can simply put the lambda after the parentheses of the function call (`map()`). For example:

```
#["one", "two"].map()[toUpperCase]
```

which further simplifies to:

```
#["one", "two"].map[toUpperCase]
```

Boxing and unboxing

Boxing refers to wrapping a primitive such as `int` or `double` into an object such as `Integer` or `Double`. Unboxing is the reverse process. The `Integer` or `Double` objects are immutable data structures necessary as many data structures assume all their contents are references to objects rather than primitives. As in `Java`, the conversion between the two representations is automatic in the vast majority of the cases. `Blang` adds boxing/unboxing to and from `IntVar` and `RealVar`,²⁷ which are mutable versions of `Integer` or `Double`. See Appendix B.3 for a discussion on why these mutable data structures are necessary in `Blang`.

Operator overloading

Operator overloading is permitted. When in the `Blang` IDE, command click on an operator to reveal its definition. One important case to be aware of is `==` which is overloaded to `.equals(...)`. For the low-level equality operator that checks if the two sides are identical (point to the same object or in the case of primitive, have the same value) use `===` (with the exception of `Double.NaN` which, following IEEE convention, is never `===` to anything).

Some useful operators that are automatically imported:

- “`0..10`”, and “`0..<11`”: These expressions are range operators and return integers 0, 1, 2, ..., 10.
- `object => lambdaExpression`: calls the lambda expression with the input given by object e.g., `new ArrayList => [add("to be added in list")]`.

When overloading operators of custom type refer to `Xtend`'s official documentation ([Xtend 2019](#)).

Parameterized types

Types can be parameterized as in `Java`'s `List` type. For example, we use `List<String>` to declare that a string will be stored, just as we would in `Java`. At the moment, models can use variables with type parameters but models themselves cannot have type parameters.

Throwing exceptions

Throw exceptions to signal abnormal behavior and to terminate the `Blang` runtime with an informative message:

```
throw new MyException("Some error message.")
```

²⁷<https://www.stat.ubc.ca/~bouchard/clang/javadoc-dsl/clang/core/IntVar.html>, <https://www.stat.ubc.ca/~bouchard/clang/javadoc-dsl/clang/core/RealVar.html>

Here `MyException` should be of type `java.lang.Throwable`. A reasonable default choice is `java.lang.RuntimeException`. To signal that the current factor has invalid parameters return the value `NEGATIVE_INFINITY`. If not possible due to a particular code structure, one can also return the value `invalidParameter`.²⁸ This will be caught and interpreted as a factor having zero probability. In contrast to Java, Blang exceptions are never required to be declared or caught. If an exception needs to be caught, the syntax is as follows:

```
try {
  // code that might throw an exception
} catch (ExceptionType exceptionName) {
  // process exception
} // optionally:
finally {
  // code executed whether the exception is thrown or not
}
```

8. Cheatsheet interlude

Learning a language can be a time-consuming task with new grammar and syntax to remember. Before continuing with more examples, ideas, and patterns, we present a condensed summary of the concepts covered thus far in the form of a recipe. We will draw connections to the GMM example from Section 6 where appropriate. For convenience, we repeat the model below:

concentration	$\alpha = [1, 1]$
proportions	$\pi \mid \alpha \sim \text{Dirichlet}(\alpha)$
labels	$z_i \mid \pi \sim \text{Categorical}(\pi)$
means	$\mu_k \sim \text{Normal}(0, 10^2)$
standard deviations	$\sigma_k \sim \text{Uniform}(0, 10)$
observations	$y_i \mid \mu, \sigma, z_i \sim \text{Normal}(\mu_{z_i}, \sigma_{z_i}^2)$

for $i \in \{1, 2, \dots, n\}$ and $k \in \{1, 2\}$.

The cheatsheet

1. Write down your package statement.

Example:

```
package jss.gmm
```

2. If it is already known which packages you will work with, then import them. We did not require additional packages for the GMM.

Example:

```
import some.other.pack
```

²⁸<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/types/StaticUtils.html>

3. Name your model.

Example:

```
model MixtureModel { ... }
```

4. Identify all model variables: observed (constant) random variables (RV), latent RVs, unknown parameters, and known (constant) parameters.

Example: observed RVs y_i , latent RVs z_i , unknown parameters π, μ_k, σ_k , and known parameter α .

5. Identify model variables' types, and values of known parameters.

Example: y_i are real numbers, z_i are integers, π is a simplex, μ_k are real numbers, σ_k are real numbers, and $\alpha = [1, 1]$.

6. Declare all observed and latent RVs, and unknown parameters with the keyword `random`; declare all known parameters with the keyword `param`.

Example:

```
random List<RealVar> observations
random Simplex pi
param Matrix concentrations
```

7. Initialize latent RVs and unknown parameters with their latent types, and initialize known parameter values. Realization of observations will be delayed until inference.

Example:

```
random List<RealVar> observations
random Simplex pi ? : latentSimplex(2)
param Matrix concentrations ? : fixedVector(1.0, 1.0)
```

8. Next we declare their respective distributions in `laws{...}`.

Example:

```
pi | concentration ~ Dirichlet(concentration)
```

9. Use `for` loops to declare over a list of variables for cleaner code.

Example:

```
for (int k : 0 ..< means.size) {
  means.get(k) ~ Normal(0.0, pow(10.0, 2.0))
}
```

10. Perform inference by using the CLI. Append `--help` to the CLI for model-specific input description.

Example:

```
blang --model jss.gmm.MixtureModel \
  --model.observations file "path/to/line_separated_values.txt"
```

9. Custom samplers for custom data structures

Although the focus of this section is on custom data structures and samplers, it will also provide insight to Blang’s underlying sampling mechanisms.

In examples we have demonstrated thus far, sampling variables could be handled via default samplers. This luxury is typically unavailable when working with complex state spaces such as trees, partitions, permutation spaces, or more generally discrete, non-ordinal spaces. In such situations Blang still assists the user in several ways described in more detail in Section 10. Here we focus on how Blang helps implement a complete sampler for a model that consists of such custom data structures.

Consider a model with latent variables taking values in a set of permutations (perfect bipartite matching). For example, *record linkage* problems (Tancredi and Liseo 2011; Steorts, Hall, and Fienberg 2016) rely on this type of latent variable. In short, record linkage is the process of matching de-identified noisy records from multiple data sources that reference the same entity or individual. For example, consider a datum taking value 170 in one dataset, and 170.1 in another dataset. The process of recognizing these two data reference the same entity is a case of record linkage. In the following, we demonstrate how to implement a custom sampler for data of type permutation.

We will implement a data type, `Permutation`, equipped with its tailor-made sampler, then

```

package jss.perm

import org.eclipse.xtend.lib.annotations.Data
import blang.mcmc.Samplers
import java.util.Random
import static java.util.Collections.sort
import static java.util.Collections.shuffle
import java.util.List

@Samplers(PermutationSampler)
@Data class Permutation {

    val List<Integer> connections

    new (int componentSize) {
        connections = (0 ..< componentSize).toList
    }

    def int componentSize() {
        return connections.size
    }

    def void sampleUniform(Random random) {
        sort(connections)
        shuffle(connections, random)
    }

    override String toString() {
        return connections.toString
    }
}

```

Figure 11: Permutation class programmed in Xtend.

apply it in the context of a model.²⁹ We begin by implementing a *class* describing how permutations will be encoded. Note a permutation can be represented with or stored as a list of integers. We present an implementation of the `Permutation` class in Figure 11, and discuss each component individually.³⁰

A first observation is the annotation `@Samplers(...)` which informs the runtime engine to sample `Permutation` objects with an instance of `PermutationSampler`.³¹ We will discuss `PermutationSampler` later.

A second observation is the annotation `@Data`.³² Briefly, this annotation should be interpreted as a *data class*, a terminology in object oriented programming (and an unfortunate clash with the conventional use of “data” in statistics), meaning the class can only declare final fields, and that `.equals`, `.hashCode` are automatically implemented in addition to other defaults.

Moving on to the main code block, as noted previously, we can represent the mathematical permutation object with a list of integers, where each element in the list is the permuted value of the element’s index. Thus we encoded the permutation object with the field `connections`, and a constructor as repeated below:

```
val List<Integer> connections

new (int componentSize) {
    connections = (0 ..< componentSize).toList
}
```

This permutation constructor is related to Section 6’s `latentSimplex(K)` constructor for simplex variables. We will see its use later to construct a latent permutation to be sampled. Technically, this is all that is required to represent the permutation type. However, it will be convenient to define a few more helper functions, in particular a function `sampleUniform` to uniformly draw a realization of a permutation.

```
def void sampleUniform(Random random) {
    sort(connections)
    shuffle(connections, random)
}
```

Notice `sampleUniform` sorts `connections`, then shuffles `connections` in-place. The sorting is required from a computational perspective to ensure the sampling is not affected by the `connection`’s current state, thus uniform when shuffled. In other words, it enforces the contract that for a given random seed encoded in the `rand` object, the behavior of the `generate` block is fully deterministic and not affected by the current state of the object. This behavior is exploited to design test cases, as in `TestCompositeModel.xtend` described shortly. The

²⁹Complete and commented implementations in this section are available in the reproduction materials located in the directory `reproduction_materials/example`, or at https://github.com/UBC-Stat-ML/JSSBlangCode/tree/master/reproduction_material/PermutationExample/src.

³⁰See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/PermutationExample/src/main/java/jss/perm/Permutation.xtend.

³¹More than one sampler can be specified as a comma-separated list, more on this in Section 12.7.

³²Complete documentation is available at <http://archive.eclipse.org/modeling/tmf/xtend/javadoc/2.9/org/eclipse/xtend/lib/Data.html>.

sampling performed in-place is a technical requirement for the inference engine, detailed in Section 12.

Finally the last piece of the puzzle is the `toString` function. Its purpose is best illustrated by an example followed by an explanation. Consider the following `permutation.csv` file below. Without overriding `toString` our sample output would read the following:

```
sample,value
0,"Permutation [
  connections = ArrayList (
    2,
    0,
    1
  )
]"
1,"Permutation [
  connections = ArrayList (
    1,
    2,
    0
  )
]"
...
```

With the `toString` function, we have:

```
sample,value
0,"[2, 0, 1]"
1,"[1, 2, 0]"
...
```

Thus we see that by overriding the default string output of our object, we enable the engine to output something more legible. One can customize this output to respect the tidy philosophy, details of which we leave to Appendix D.

With all the pieces in place for our `Permutation` class, we are now ready to discuss samplers. To perform posterior inference on permutation spaces, we need an invariant sampler designed specifically for the object `Permutation`. In this example, we assume familiarity with the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953), and begin by presenting `PermutationSampler.xtend` in Figure 12,³³ followed by a breakdown of its main components.

A first observation is the *implementation* of the `Sampler interface`, and the two annotations `@SampledVariable` and `@ConnectedFactor`:³⁴

```
class PermutationSampler implements Sampler {
  @SampledVariable Permutation permutation
  @ConnectedFactor List<LogScaleFactor> numericFactors
```

³³See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/PermutationExample/src/main/java/jss/perm/PermutationSampler.xtend.

³⁴For more on interfaces, see <https://docs.oracle.com/javase/tutorial/java/concepts/interface.html>.

```

package jss.perm

import java.util.List
import bayonet.distributions.Random
import blang.core.LogScaleFactor
import blang.mcmc.ConnectedFactor
import blang.mcmc.SampledVariable
import blang.mcmc.Sampler
import blang.distributions.Generators
import static java.lang.Math.exp
import static java.lang.Math.min
import static extension java.util.Collections.swap

class PermutationSampler implements Sampler {
    @SampledVariable Permutation permutation
    @ConnectedFactor List<LogScaleFactor> numericFactors

    override void execute(Random rand) {
        val n = permutation.componentSize
        val i = Generators.discreteUniform(rand, 0, n)
        val j = Generators.discreteUniform(rand, 0, n)

        val currentLogDensity = logDensity()
        permutation.connections.swap(i,j)
        val newLogDensity = logDensity()

        val acceptProb = min(1.0, exp(newLogDensity - currentLogDensity))
        val accept = Generators.bernoulli(rand, acceptProb)
        if (!accept) {
            permutation.connections.swap(i, j)
        }
    }
    def double logDensity() {
        var double sum=0.0
        for (LogScaleFactor f : numericFactors) sum += f.logDensity()
        return sum
    }
}

```

Figure 12: Permutation sampler programmed in Xtend.

Briefly, this implies the `PermutationSampler` class necessarily implements methods specified in the interface `Sampler`, namely `execute`. The `execute` method is invoked with each iteration of the inference algorithm (Section 12.1), and updates our variable of interest in-place. The field annotated with `@SampledVariable` will automatically be populated with an instance of the object to be sampled, in this example, an instance of `Permutation`. This annotation in tandem with `@Samplers` enables the linkage of variables and samplers.³⁵ Similarly, the field annotated with `@ConnectedFactor` will automatically be populated with factors dependent on the sampled object, which is inferred automatically via a factor graph built from scope analysis (described in detail in Section 12). By default once a type and its sampler have been implemented, variables of such type will be sampled with this sampler. We discussed how this default is altered in Section 7.10, and will discuss it again in Section 12.7.

With this setup, we are ready to implement the Metropolis algorithm for permutations as

³⁵Sampling of multiple variables can also be performed. For example, the SDK incorporates an elliptic slice algorithm which samples many real variables at once, see <https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/distributions/NormalField.bl> and <https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/mcmc/EllipticalSliceSampler.xtend>.

```

override void execute(Random rand) {
    val n = permutation.componentSize
    val i = rand.nextInt(n)
    val j = rand.nextInt(n)

    val currentLogDensity = logDensity()
    permutation.connections.swap(i,j)
    val newLogDensity = logDensity()

    val acceptProb = min(1.0, exp(newLogDensity - currentLogDensity))
    val accept = Generators.bernoulli(rand, acceptProb)
    if (!accept) {
        permutation.connections.swap(i, j)
    }
}

def double logDensity() {
    var double sum=0.0
    for (LogScaleFactor f : numericFactors) sum += f.logDensity()
    return sum
}

```

Figure 13: Implementation of a Metropolis algorithm for permutations.

shown in Figure 13. The implementation of `execute()` is a standard Metropolis algorithm (Metropolis *et al.* 1953), which invokes `logDensity` when density evaluation is required. Since the field `numericFactors` is a list of all log factors dependent on our variable, the `logDensity` method merely returns the sum of log factors.

Notice the syntax `Generators.bernoulli(rand, acceptProb)` is used to determine acceptance of the proposal. This syntax is equivalent to `Bernoulli.distribution(p).sample(rand)` (in fact the latter calls the former). However the second variant creates an intermediate object of type `IntDistribution` which could be a performance issue as the body of the sampling algorithm is in the inner loop of inference. In other contexts, having this intermediate object is useful, e.g., if one would like to provide a distribution as input parameter to another distribution, as in Section 11.4. As for the relationship with `... ~ Bernoulli(...)`, recall that the difference is that `... ~ Bernoulli(...)` is used in a declarative context, while the first two syntaxes are for imperative blocks such as MCMC samplers (of course, in all three variants there is no code duplications in the SDK, i.e., higher-level functions such as the declarative syntax call lower level implementations).

With our custom `Permutation` type and `PermutationSampler` in place, we are ready to apply them in models. Figure 14 shows an example of a uniform distribution over the permutation space as implemented in `UniformPermutation.bl`.³⁶

As we have seen before, the `logf` block provides a method for evaluating log densities, while the `generate` block provides a method for sampling permutations in place. In this case, `logf` returns the log density of a permutation with uniform distribution, and `generate` samples a permutation uniformly. As for `logFactorial`, which computes $\log(n!)$, it is part of the automatically imported functions described in Section 7.6 (a list of the most commonly used automatically imported functions can also be found in Appendix F).

³⁶See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/PermutationExample/src/main/java/jss/perm/UniformPermutation.bl.

```

model UniformPermutation {
  random Permutation permutation

  laws {
    logf(permutation) {
      - logFactorial(permutation.componentSize)
    }
  }

  generate(rand) {
    permutation.sampleUniform(rand)
  }
}

```

Figure 14: Uniform distribution over a permutation space programmed in Blang, `Uniform.bl`.

```

package jss.perm

model CompositeModel {
  random List<RealVar> y ?: fixedReallist(2.1,-0.3,0.8)
  random Permutation permutation ?: new Permutation(y.size)

  laws {
    permutation ~ UniformPermutation
    for (int i : 0 ..< y.size) {
      y.get(i) | permutation, i
      ~ Normal(permutation.getConnections.get(i), 0.3)
    }
  }
}

```

Figure 15: Blang model illustrating use of composite laws with `UniformPermutation` sampler.

As with any distribution, model `UniformPermutation` can be used in composition with other models. Here we present a minimal, illustrative example in Figure 15.³⁷

This should look rather similar to other models, with the exception of the use of a custom constructor `new Permutation(y.size)` to instantiate the latent permutation variable. `CompositeModel` provides a toy example of how one can incorporate `UniformPermutation` into larger models. An example of custom data types with emphasis on a practical application using a spike and slab model (Mitchell and Beauchamp 1988) is presented in Appendix A.2. This concludes our tutorial on creating custom data types and samplers. We dedicate the remainder of this section to showcasing some available resources that assist users in testing the correctness of samplers.

A first test utility provided by the SDK is `DiscreteMCTest`, which is specialized to fully-discrete spaces. The idea behind `DiscreteMCTest` is that, for small discrete spaces, we can explicitly form a sparse transition matrix and numerically check properties such as invariance and irreducibility. In our experience, many software defects can be found in problems just large enough to achieve code coverage.

To run tests, we need to setup a project directory, i.e., `create-blang-gradle-project`

³⁷See https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/PermutationExample/src/main/java/jss/perm/CompositeModel.bl.

`--name PermutationExample`.³⁸ This will create a directory named “PermutationExample”, with directory structure `src/main/java`. Place our implementations in this directory, and create the additional directory `PermutationExample/src/test/java/`. Making sure the package names are matching, place the file `TestCompositeModel.xtend` in the testing directory `src/test/java/jss/perm/` with implementations as in Figure 16. The test can be performed using `./gradlew test` while in the `PermutationExample` directory, or via the Eclipse IDE.

The `DiscreteMCTest` object takes in two arguments. The first argument is a small, discrete model. The construction of this discrete model is achieved using the compiled builder in `CompositeModel.java` (or in general, `ModelName.java`). The second argument is a lambda function (denoted by square brackets) that accepts a model and creates a new object encoding the identity of the current configuration, with identity being mediated by the `.equals()` function of the returned object.

As `DiscreteMCTest` is created (i.e., handled in the construction of the object), the samplers involved in the input model are automatically translated into explicit sparse transition matrices, via a type of non-standard evaluation of the sampling code.³⁹ Given these inputs, irreducibility and invariance tests boil down to an application of linear algebra and graph algorithms.

Detailed testing resources are discussed in Section 10.5, including tests for models defined on continuous spaces.

10. Tools and software development kit

Blang comes with “batteries included”: more than just a language, it is a suite of tools and libraries supporting common tasks in Bayesian data analysis. In this section, we present an overview of these libraries. Briefly, we start with a description of the Blang integrated development environment (IDE), followed by a discussion on how input of data is handled in Blang. This includes an introduction to implementing plate and plated variables for plate notation used in traditional graphical models. Next, we discuss how samplers, distributions, and other components fit into the core inference algorithms’ architecture. Finally, we conclude with brief discussions on post-processing options, monitoring logs, testing frameworks, and additional packages and dependencies.

10.1. Integrated development environment (IDE)

Integrated development environments are software applications built for software construction. They are typically equipped with features such as syntax highlighting, code completion, refactoring, debugging and other tools that assist programmers in software development.

Desktop IDE

We provide an IDE for Blang built on Eclipse. The only requirement is that Java 11, 13, or 15

³⁸Commented implementations on testing are available in the reproduction materials located in the directory `PermutationExample`.

³⁹More information is available at https://www.stat.ubc.ca/~bouchard/blang/Testing_Blang_models.html under “Exhaustive tests.”

```

package jss.perm

import static blang.types.StaticUtils.*
import blang.runtime.SampledModel
import blang.validation.DiscreteMCTest
import com.rits.cloning.Cloner
import org.junit.Test
import static org.apache.commons.math3.util.CombinatoricsUtils.factorial
import static java.lang.Math.pow
import blang.runtime.internals.objectgraph.GraphAnalysis
import blang.runtime.Observations
import blang.types.ExtensionUtils

class TestCompositeModel {

    val static y = fixedRealList(2.1, -0.3, 0.8)
    val static CompositeModel compositeModel = new CompositeModel.Builder()
        .setY(y)
        .setPermutation(new Permutation(y.size))
        .build

    val static observations = {
        val Observations result = new Observations
        result.markAsObserved(y)
        result
    }

    val static DiscreteMCTest test =
        new DiscreteMCTest(
            new SampledModel(new GraphAnalysis(compositeModel, observations)),
            [
                val CompositeModel cm = model as CompositeModel
                return new Cloner().deepClone(cm.permutation)
            ]
        )

    @Test
    def void stateSize() {
        test.verbose = true
        test.checkStateSpaceSize(factorial(y.size) as int)
    }

    @Test
    def void invariance() {
        test.verbose = true
        test.checkInvariance
    }

    @Test
    def void irreducibility() {
        test.verbose = true
        test.checkIrreducibility
    }
}

```

Figure 16: Xtend implementation to test a composite model `TestCompositeModel.xtend`.

should be installed. There are two ways to install it: one using the pre-packaged version, the other by adding a plug-in to an already installed Eclipse instance. The former method is more straightforward but currently we only distribute the pre-packaged Blang Eclipse for Mac OS X (tested with Mac OS 10.11.6, 10.14.6, 10.15.7). The latter method supports Mac OS X, Windows and Linux.

For the Mac-specific method, download the IDE available at <https://www.stat.ubc.ca/~bouchard/blang/downloads/blang-mac-4.0.7.zip>. Unzip the downloaded file and copy the contents to a directory of your choice. The folder contains both the IDE, a template for your own projects, and some command line tools. The first time you try to launch Blang IDE, depending on the version of Mac OS X and/or security settings, you may get a message saying the “app is not registered with Apple by an identified developer.” To work around this, follow the instructions (from Apple) the first time you open the Blang IDE (then Mac OS will remember your decision for subsequent launches) given at <https://support.apple.com/en-ca/guide/mac-help/mh40616/mac>.

The second installation process for the Blang IDE, which is the most portable across platforms, is the following:

1. Install *DSL tools for Eclipse*, which can be downloaded from the Eclipse website.⁴⁰ At the time of writing, the supported version is Eclipse IDE for Java and DSL (domain specific language) Developers, Release 2020-12 R.⁴¹ Using the standard version of Eclipse (i.e., not the DSL version) and/or a different version is unlikely to work.
2. From Eclipse: select *Install New Software* from the *Help* menu.
3. Click *Add* and enter <https://www.stat.ubc.ca/~bouchard/maven/blang-eclipse-plugin-4.0.7/> in the location field.
4. Click *Select All, Next*, then follow instructions as prompted.

To create or open a new project, follow these instructions:

1. (Skip this step if you want to open an existing project.) To create a template for a new project, if you have the Blang CLI installed, type `create-blang-gradle-project --githubOrganization myOrg --name myProject` where you should replace “myOrg” by the name of your organization, and “myProject” by the required name for the project. You can also find a template project at <https://github.com/UBC-Stat-ML/blangExample> (the method based on `create-blang-gradle-project` as it guarantees the library versions will be in sync with the version of Blang used by the CLI).
2. The next step is to generate configuration files suitable for Eclipse. This can be done by using the command `bash setup-eclipse.sh` which can be found at the root of a freshly generated project, or `./gradlew assemble eclipse` if an older template project is used.
3. Starting from Eclipse’s menus, select *File > Import > General > Existing project into Workspace*. Select the root of the project you created in the previous step.
4. The Blang project is ready. In the left tool bar in Eclipse, the project is in the file explorer. Right click on `src/main/java/[package name]/` and select the contextual menu *New > File*. Name the file `MyModel.bl`. The extension choice must always be `.bl`.

⁴⁰<https://www.eclipse.org/>

⁴¹Linked at <https://www.eclipse.org/downloads/packages/release/2020-12/r/eclipse-ide-java-and-dsl-developers>.

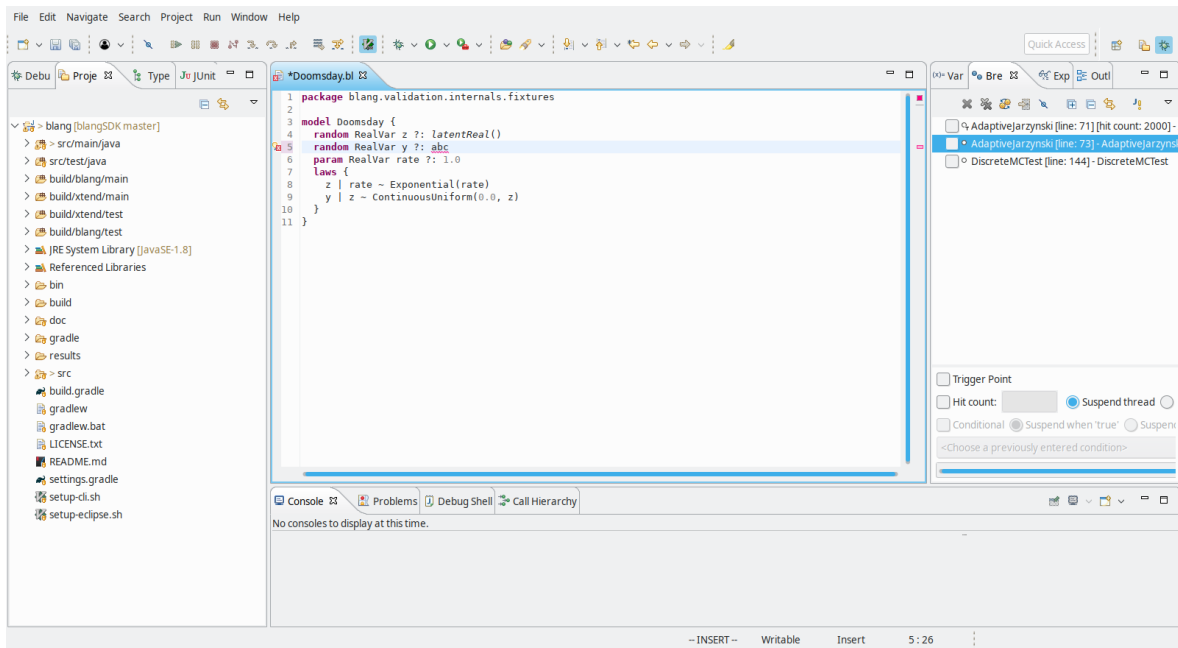


Figure 17: A preview of the Blang IDE. A warning of a syntactical error is underlined in red.

The key IDE features useful for development include:

- The ability to navigate Blang code by, for example, holding the “command” key while clicking on any Blang symbol to jump to its definition, or hovering on the symbol to view its documentation. This and other related features are possible thanks to the static type system used by Blang.
- Incremental compilation in parallel in the background, which implies little time is spent waiting for compilation on modern multicore architectures. It also means that error messages appear interactively as the user types. See example in Figure 17.
- Quickly viewing the generated Java files, by right clicking anywhere in a Blang editor and selecting “Open Generated File”.
- From any generated file, inference on the model can be launched by right-clicking “Run As... Java Application”. After doing this the first time, a shortcut is accessible via the menu “Run > Run Configurations...” Run Configurations allow setting the command-line arguments being passed to Blang. When using this method to launch a Blang execution, note that the argument `--model [name of model]` should be skipped.
- A full-feature debugger is built-in. Double clicking on the left margin of a Blang or Xtend file sets a break point. Use the menu “Debug > Debug Configurations” to start the debugger.
- Being built on Eclipse, the IDE also inherits Eclipse’s comprehensive set of features, such as utilities for unit testing, code coverage analysis, git integration, visualization of call and type hierarchies among others.

More information on the Blang IDE is available from the Blang documentation page, https://www.stat.ubc.ca/~bouchard/clang/Blang_IDE.html.

Web IDE

To facilitate deployment on large number of cores on the cloud, for example in a teaching or reproducible research context, Blang is also available on the Web scientific platform Silico (<https://silico.io/>).

To setup a Blang project in Silico, create a `Model` from the user profile page, and create a file with `.bl` extension. Command-line arguments can be passed in by pasting them in a file called `configuration.txt`.

10.2. Data types provided in the SDK

The interfaces `RealVar` and `IntVar` are automatically imported.⁴² They can be either latent (unobserved, sampled), or fixed (conditioned upon). See Table 10 in Appendix F for commonly used functions to provide default initializations to these basic random variables.

Blang's linear algebra is based on `xlinear` (for more information see Appendix C.3) which is in turn based on a portfolio of established libraries. The basic classes available are `Matrix`, `DenseMatrix`, and `SparseMatrix`. Blang/XBase allows operator overloading (Efttinge *et al.* 2013), so it is possible to write expressions of the type `matrix1 * matrix2`, `2.0 * matrix`, and so on. Vectors do not have a distinct type, they are just $1 \times n$ or $n \times 1$ matrices. Standard operations are supported using unsurprising syntax, e.g., `identity(100_000)` (underscore delimited 100,000), `ones(3, 3)`, `matrix.norm`, `matrix.sum`, `matrix.readOnlyView`, `matrix.slice(1, 3, 0, 2)`, `matrix.cholesky`, etc.⁴³

Blang augments `xlinear` with two specialized types of matrices: `Simplex`, vector of positive numbers summing to one, and `TransitionMatrix`. Refer to Table 11 in Appendix F for key functions related to these specialized types of matrices.

10.3. Distributions

A range of distributions are included in the SDK. See Appendix E for the current list. These distributions are themselves written in Blang. The SDK also contains tests covering all the included distributions. Our development workflow performs all the unit tests each time a commit is made in the Blang GitHub repository.

The implementation of the random number generators used in forward simulation of the SDK distributions are all grouped in the file `Generators`.⁴⁴

10.4. Input

Inputs are parsed and managed via the `inits` package's injection framework. Model variables can be provided a default initialization in the model's `.bl` file, or they can be initialized

⁴²<https://www.stat.ubc.ca/~bouchard/clang/javadoc-dsl/clang/core/IntVar.html>, <https://www.stat.ubc.ca/~bouchard/clang/javadoc-dsl/clang/core/RealVar.html>

⁴³See <https://github.com/UBC-Stat-ML/xlinear> for more information on `xlinear`.

⁴⁴<https://www.stat.ubc.ca/~bouchard/clang/javadoc-sdk/clang/distributions/Generators.html>

```

import blang.inits.ConstructorArg;
import blang.inits.DesignatedConstructor;
import blang.inits.GlobalArg;
import blang.runtime.Observations;

class Type1
{
    ...
    @DesignatedConstructor
    def static Type1 loadObservedData(
        @ConstructorArg(value = "file") File file,
        @ConstructorArg(value = "option") Integer x,
        @GlobalArg Observations observations)
    {
        val Type1 result = doSomethingWith(file, x) // Parse the file
        observations.markAsObserved(result)
        return result
    }
    ...
}

```

Figure 18: Xtend code illustrating how to parse custom data types `Type1.xtend`.

with arguments through the CLI. Should both methods exist, the latter takes precedence; an example exposing only the pertinent snippets of code is shown below in `Name.bl`:

```

model Name {
    param IntVar h ? : 3
    param IntVar a
    random Type1 p
    ...
}

```

The field `h` is initialized to 3 by default, but can be overridden by the command-line argument `to`, say 5, `--model.h 5`. The field `a` must be assigned a value, say 9, via the CLI through `--model.a 9`. For observations or custom data types such as `Type1`, annotations can be easily added to control parsing. The following file `Type1.xtend` is an example of a constructor that can parse command-line arguments such as `--model.p.file abc.csv --model.p.option 2`.

In the code above, `doSomethingWith(file, x)` is any user-defined function that returns the parsed result as desired.

Additional information is provided at the relevant Blang documentation pages.⁴⁵

Plate notation

Simple collections of random variables can be handled via Java's built-in `List` and related data structures. However, this can quickly become cumbersome and error-prone when working with sophisticated hierarchical Bayesian models. To address this problem, Blang provides a specialized data structure based on the *plate notation*, a method for representing repeated random variables in a graphical model. A concise encoding of these models can be achieved with built-in types `Plate` and `Plated`.

⁴⁵<https://www.stat.ubc.ca/~bouchard/blang/Javadoc.html>.

Country	Rocket	nLaunches	nFails
CHN	Chang Zheng 1	2	0
ESA	Ariane 44P	15	0
RUS	Molniya 8K78M	272	13
USA	Delta 2914	30	2
⋮	⋮	⋮	⋮

Table 3: Rocket launching data in tidy format.

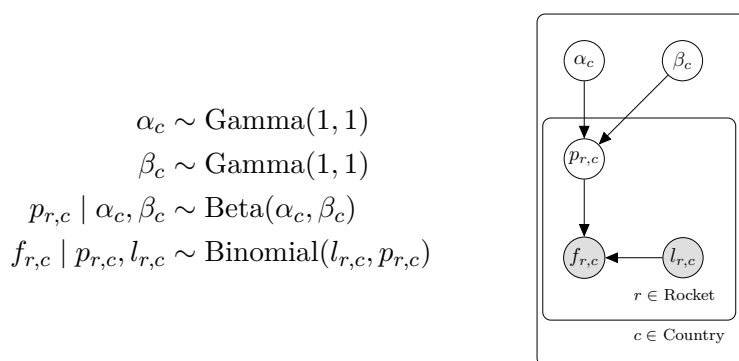


Figure 19: Left: a toy hierarchical model for the rocket launching data set. Indices c, r index countries and rockets respectively. Observations $f_{r,c}, l_{r,c}$ are the number of failures and launches for rocket r in country c respectively. Latent variables $p_{r,c}, \alpha_c, \beta_c$ are parameters of interest. Right: A graphical representation of the hierarchical rocket model with plate notation.

Consider the rocket launching data in tidy format in Table 3. “Country” is the origin of the rocket, “Rocket” is the name of rockets, “nLaunches” is the number of launches, and “nFails” is the number of failed launches. A toy model for this data set is the hierarchical model in Figure 19. The Blang code for the rocket model is provided in Figure 20.⁴⁶ The complete code and data can also be found in the prepackaged repository of examples.

A first observation is the additional param `GlobalDataSource data`, which we have not seen in our previous models. We will discuss its function in more details shortly. At a high level, it is used to specify a CSV file from which many variables will be parsed.

A second observation is the use of `Plate<...>` and `Plated<...>` types.

```
param Plate<String> countries
param Plate<String> rockets
```

⁴⁶https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/hier/Rocket.bl

```

package jss.hier

model Rocket {

    param GlobalDataSource data

    param Plate<String> countries
    param Plate<String> rockets

    random Plated<RealVar> alpha
    random Plated<RealVar> beta
    random Plated<RealVar> prob
    random Plated<IntVar> nFails
    param Plated<Integer> nLaunches

    laws {
        for(Index<String> c : countries.indices()){
            alpha.get(c) ~ Gamma(1,1)
            beta.get(c) ~ Gamma(1,1)

            for(Index<String> r : rockets.indices(c)){
                prob.get(r, c) |
                RealVar a = alpha.get(c),
                RealVar b = beta.get(c)
                ~ Beta(a, b)
                nFails.get(c, r) |
                RealVar p = prob.get(r, c),
                Integer n = nLaunches.get(r, c)
                ~ Binomial(n, p)
            }
        }
    }
}

```

Figure 20: Blang code for the rocket model `Rocket.bl`.

```

random Plated<RealVar> alpha
random Plated<RealVar> beta
random Plated<RealVar> prob
random Plated<IntVar> nFails
param Plated<IntVar> nLaunches

```

A `Plate` is a collection of indices, such as country and rocket indices (columns one and two of our example data set above). As they are non-random known indices, we declare plates with `param`. On the other hand `Plated` types, as its name suggests, are variables within plates. The usual rules for selection between `param` or `random` apply to plated variables (see Section 5.1).

With this setup, we are ready to examine the `laws` block.

```

for(Index<String> c : countries.indices()){
    alpha.get(c) ~ Gamma(1,1)
    beta.get(c) ~ Gamma(1,1)

    for(Index<String> r : rockets.indices(c)){
        prob.get(r, c) |
        RealVar a = alpha.get(c),

```

```

    RealVar b = beta.get(c)
    ~ Beta(a, b)
  nFails.get(c, r) |
    RealVar p = prob.get(r, c),
    Integer n = nLaunches.get(r, c)
    ~ Binomial(n, p)
  }
}

```

This should look rather similar to code we have presented thus far. We highlight the key differences: first, the set of index values is obtained by appending `.indices` to a `Plate` variable. Each index is of type `Index<T>`, where `T` is the same type as the corresponding `Plate<T>`. Plated variables can subsequently be retrieved by using `.get()`. Second, notice the syntax of the second `for` loop over rocket indices, in particular `rockets.indices(c)`. This syntax retrieves the set of rocket indices such that its country index is `c`. Lastly, we note the ordering of indices within `.get()` is exchangeable, for example, `nFails.get(c, r)` is equivalent to `nFails.get(r, c)`. This is possible since `Index<...>` objects keep track of which plate they belong to. Additional methods available for types `Index<>` are described in Table 13 in Appendix F.

With variables, parameters, and laws declared, we tie these concepts back to the promised discussion of `param GlobalDataSource data`. Its purpose becomes clear when we invoke `blang` and its corresponding arguments:

```

$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example/
$ blang --model jss.hier.Rocket \
> --model.data data/rockets.csv \
> --model.countries.name Country \
> --model.rockets.name Rocket \

```

Variables of type `Plate` and `Plated` can be put in correspondence with a column in a tidy CSV file. Command-line arguments can be used to set the CSV file for *each* variable individually. *Alternatively* by declaring a dummy variable of type `GlobalDataSource`, here called `data`, we can set a default CSV file that will be used by default by all `Plate` and `Plated` variables. In our example, specifying the default CSV is achieved via `--model.data pathToData/data.csv`. For each `Plate` and `Plated` variable, the data input algorithm will attempt to find a column in the CSV file matching with the variable name. The algorithm does not require all columns in the CSV file to be matched to `Plate` or `Plated` variables.

By default, matching is done by using the same string for the column header as the variable name, but this can be overridden via CLI arguments. In our example, this is achieved via e.g., `--model.rockets.name Rocket`. Notice we did not require this argument for `nFails`, as the column name in the CSV file is also `nFails`.

When a plated variable is not found in the CSV file, it is assumed to be latent (a message is displayed to standard out when this happens). Should a plate not correspond to a column in the CSV, then its `maxSize` should be set via, e.g., `--model.varName.maxSize 3`, or initialized in the model. An example using the CLI is presented in the advanced tutorial in

```

package jss.gmm

model MixtureModelPlated {

  param GlobalDataSource data

  param Integer          K   ?: 2
  param Plate<Integer>  N
  param Plate<Integer>  components   ?: Plate.ofIntegers("components", K)

  random Plated<IntVar>  z
  random Plated<RealVar> y
  random Plated<RealVar> mu
  random Plated<RealVar> sd
  random Simplex        pi   ?: latentSimplex(K)

  laws {
    pi | K ~ SymmetricDirichlet(K, 1.0)

    for (Index<Integer> k : components.indices) {
      mu.get(k) ~ Normal(0.0, 100.0)
      sd.get(k) ~ ContinuousUniform(0.0, 10.0)
    }

    for (Index<Integer> i : N.indices) {
      z.get(i) | pi ~ Categorical(pi)
      y.get(i) | List<RealVar> muList = mu.asList(components),
                List<RealVar> sdList = sd.asList(components),
                IntVar k = z.get(i)
                ~ Normal(muList.get(k), pow(sdList.get(k), 2.0))
    }
  }
}

```

Figure 21: Implementation of a mixture model using plate syntax, `MixtureModelPlated.bl`.

Appendix A.2,⁴⁷ and an example using default initializations is shown below.

Recall the Gaussian mixture model from Section 6. We provide an implementation of the same model with plate syntax in Figure 21.⁴⁸ A first observation is the use of `Plate.ofIntegers()` to initialize a plate with a predetermined size. The function `ofIntegers()` takes in two arguments: a column name, and a maximum size. For other related functions, see Appendix F. A second observation is the `asList()` function, which returns the given plate (`components`) as a list.

```

y.get(i) | List<RealVar> muList = mu.asList(components),
          List<RealVar> sdList = sd.asList(components),
          IntVar k = z.get(i)
          ~ Normal(muList.get(k), pow(sdList.get(k), 2.0))

```

Similar conversion utilities automatically imported from `ExtensionUtils` are documented in Table 12 of Appendix F.⁴⁹ With this setup, our command line arguments for inference can be written succinctly:

⁴⁷Use the argument `--model Rocket --help` for full documentation.

⁴⁸https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/gmm/MixtureModelPlated.bl

⁴⁹<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/types/ExtensionUtils.html>

```

model PlatedMatrixExample {
  param Plate<String> dims
  param Plate<String> replicates
  random PlatedMatrix vectors
  laws {
    for (Index<String> n : replicates.indices) {
      vectors.getDenseVector(dims, n) |
      int size = dims.indices.size
      ~ MultivariateNormal(dense(size), identity(size).cholesky)
    }
  }
}

```

Figure 22: Implementation of plated matrix in Blang, `PlatedMatrixExample.bl`.

```

$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example/
$ blang --model jss.gmm.MixtureModelPlated \
> --model.data data/obs1Plated.csv

```

PlatedMatrix

One special case of a `Plated` variable is the type `PlatedMatrix`,⁵⁰ which is a built-in type that facilitates easy representation of higher dimensional random variables such as random vectors or matrices, as well as lists and arrays of vectors and matrices.

`PlatedMatrix` can be used to represent both random vectors and matrices that are enclosed within a `Plate`. `PlatedMatrix` generally works in the same way as `Plated`, but provides specialized mechanisms to access vectors, matrices, simplices, etc. For example, to access a dense vector, use the method `myPlatedMatrix.getDenseVector(myRowPlate, myParentIndex1, myParentIndex2, ...)`. Here `myRowPlate` refers to the plate from which row indices will be constructed. The indices can be of arbitrary type (e.g., integer or string-valued); a mapping of indices to non-negative integers is maintained internally.

The other arguments, `myParentIndex1`, `myParentIndex2`, ... correspond to the indices for the plates in which this vector belongs to. For example, see how a set of vectors can be obtained in the Blang file `PlatedMatrixExample.bl` given also in Figure 22.

10.5. Testing framework

There is considerable emphasis in the Markov chain Monte Carlo (MCMC) literature on efficiency, but much less on correctness, in the sense of the implementation being ergodic with respect to the distribution of interest. Consider a Monte Carlo procedure producing samples X_1, X_2, \dots , targeting some distribution π . We say the procedure is correct if its ergodic averages for any integrable function f admits a law of large numbers converging to the posterior expectation of f under the target π ,

$$\frac{1}{N} \sum_{i=1}^N f(X_i) \rightarrow \int f(x)\pi(x)dx,$$

almost surely.

⁵⁰<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/types/PlatedMatrix.html>

```

package blang.validation

import java.util.function.Supplier
import bayonet.distributions.ExhaustiveDebugRandom

class UnbiasednessTest {
    def static double expectedZEstimate(Supplier<Double> logZEstimator,
                                       ExhaustiveDebugRandom exhaustiveRand) {
        var expectation = 0.0
        var nProgramTraces = 0
        while (exhaustiveRand.hasNext) {
            val logZ = logZEstimator.get
            expectation += Math.exp(logZ) * exhaustiveRand.lastProbability
            nProgramTraces++
        }
        println("nProgramTraces = " + nProgramTraces)
        return expectation
    }
}

```

Figure 23: Implementation of the unbiasedness test, `UnbiasednessTest.xtend`.

Two common defects of an incorrect Monte Carlo procedure include erroneous mathematical derivations of algorithms and software implementation bugs. This section discusses the tools available to test and detect both types of problems.

Exhaustive random objects

We provide in `bayonet.distributions.ExhaustiveDebugRandom` a non-standard replacement implementation of `bayonet.distributions.Random` which enumerates all the possible realizations of an arbitrary finite random process along with the probability of each realization.

Testing unbiasedness

We use `ExhaustiveRandom` to test the unbiasedness of the normalization constant estimate provided by our sequential Monte Carlo (SMC) implementation. The code forming the basis of this test is shown in Figure 23. The code defines a function, `expectedZEstimate` which takes as input an estimator `logZEstimator` and an `ExhaustiveRandom` object. The estimator is assumed to use internally the `ExhaustiveRandom` object to provide a randomized estimate. Assuming that the estimator is defined on a finite probability space, the code can therefore compute the exact value of the expectation of `logZEstimator`.

Note that SMC executed on a small finite model, e.g., a short hidden Markov model, has a finite number of possible execution traces (defined as all possible intermediate particles, i.e., possible proposal and resampling vectors).

This code is called in our continuous integration test suite to verify unbiasedness of our SMC implementation on a model small enough for the number of possible execution traces to be manageable while achieving code coverage.⁵¹

The output of a test based on the unbiasedness test has the form:

⁵¹<https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/test/java/blang/TestSMCUnbiasness.xtend>

```
nProgramTraces = 23868
true normalization constant Z: 0.345
expected Z estimate over all traces: 0.34500000000000164
```

where `true normalization constant Z` is computed by explicitly enumerating all states in the space of the posterior (SMC code is not needed since the state space is finite), and `expected Z estimate over all traces` is computed by enumerating all SMC execution traces along with their probabilities.

Tests based on linear algebra

We can leverage the `ExhaustiveRandom` object to assert the invariance and irreducibility of a transition kernel, in a similar flavor to the test for unbiasedness. As such, the aim of this short section is to highlight the key ideas of such tests, with details deferred to reference code. The `DiscreteMCTest` contains algorithms that use `ExhaustiveDebugRandom` to check via linear algebra whether Markov kernels on small discrete (finite state space) models are invariant and irreducible.⁵²

Algorithmically, `DiscreteMCTest` takes a model and a kernel, and constructs the corresponding sparse transition matrix. From this matrix it is then trivial to check, numerically, irreducibility and invariance via linear algebra and graph algorithms. See `TestDiscreteModels` for an example.⁵³

Exact invariance test

Tests discussed thus far focused around the idea of exhaustively enumerating outcomes and probabilities. Although these tests are attractive due to their deterministic property, they are only applicable to a small set of models, namely models with finite state spaces. These tests are not applicable to models with continuous state spaces.

For continuous models, we provide a modified form of the Geweke test (Geweke 2004), which we call the exact invariance test. Consider the goal of testing the invariance of a kernel T with respect to some target distribution $\pi(\theta | y) \propto p_\theta(\theta)p_{y|\theta}(y | \theta)$. Assume T is a combination of individual kernels T_i for $i = 1, 2, \dots, Q$. Briefly, the Geweke test examines the correctness of an MCMC procedure by comparing two sets of simulated random variables, $F = \{F_1, F_2, \dots, F_{M_1}\}$ and $G = \{G_1, G_2, \dots, G_{M_2}\}$ using an approximate test based on an asymptotic result.

The set F is generated by the marginal-conditional simulator defined by iterating the three steps:

1. $\theta_m \sim p_\theta(\cdot)$,
2. $y_m | \theta_m \sim p_{y|\theta}(\cdot | \theta_m)$,
3. $F_m = f(\theta_m, y_m)$,

for $m = 1, 2, \dots, M_1$ and some integrable, real-valued test function f . Similarly, G is generated by the successive-conditional simulator defined by the following steps:

⁵²<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/validation/DiscreteMCTest.html>

⁵³<https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/test/java/blang/>

1. $\theta_1 \sim p_\theta(\cdot)$,
2. $y_1 \mid \theta_1 \sim p_{y|\theta}(\cdot \mid \theta_1)$,
3. $G_1 = f(\theta_1, y_1)$,
4. iterate for $m = 2, \dots, M_2$:
 - (a) $\theta_m \mid \theta_{m-1}, y_{m-1} \sim T(\cdot \mid \theta_{m-1}, y_{m-1})$
 - (b) $y_m \mid \theta_m \sim p_{y|\theta}(\cdot \mid \theta_m)$
 - (c) $G_m = f(\theta_m, y_m)$.

The Geweke test is based on the observation that both F and G can be used to approximate the expectation of f under the joint distribution of p_θ and $p_{y|\theta}$. However, there are several limitations to the original Geweke approach:

1. The validity of the approximate test relies on T being irreducible. As a consequence, individual kernels T_i 's often cannot be tested in isolation.
2. The test is an approximate test relying on asymptotics. It is difficult to verify the accuracy of this asymptotic result in practice. Furthermore, the problem is compounded when several such tests need to be combined using a multiple-testing framework.
3. The validity of the approximate test also relies on a central limit theorem for Markov chains to hold, which typically involves establishing geometric ergodicity. The task of proving geometric ergodicity is model-dependent and rather involved.

To address the problems above, Blang employs a modified version of the Geweke test, the exact invariance test (EIT). The EIT does not rely on irreducibility of T_i 's, thus allowing individual tests. Furthermore, the test does not rely on establishing geometric ergodicity, and as its name suggests, it is an exact test independent of asymptotics.

Similar to the Geweke test, it compares two sets of samples F and $H = \{H_1, H_2, \dots, H_{M_3}\}$. The samples H are generated from the exact invariant simulator defined by the steps:

1. $\theta_{1,m} \sim p_\theta(\cdot)$,
2. $y_{1,m} \mid \theta_{1,m} \sim p_{y|\theta}(\cdot \mid \theta_{1,m})$
3. For $k = 2, 3, \dots, K$
 - (a) $\theta_{k,m} \mid \theta_{k-1,m}, y_{k-1,m} \sim T_i(\cdot \mid \theta_{k-1,m}, y_{k-1,m})$
4. $H_m = f(\theta_{K,m}, y_{K,m})$.

By construction, for any $K \geq 1$, $j \in \{1, 2, \dots, M_1\}$, $l \in \{1, 2, \dots, M_3\}$, F_j and H_l are equal in distribution if and only if T_i is π -invariant. Thus the appropriate exact tests (e.g., Fisher's exact test), or well-understood asymptotic tests (e.g., Kolmogorov-Smirnov) may be employed. Note H_m 's are also independent, thus the asymptotics do not rely on irreducibility nor geometric ergodicity; standard tests for independent and identically distributed (IID) random variables can be employed.

An example of how EIT is used in Blang to automatically test all distributions in the SDK can be found in <https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/test/java/blang/TestSDKDistributions.xtend>. A complete example of an EIT for our permutation model (Section 9) is provided in the reproduction materials.

10.6. Package distribution and injection

Distributing and reusing packages is standard practice in software development. Any user can create a model and publish it in a versioned fashion via GitHub.⁵⁴

To use a package developed by another user, Blang projects compiled via the CLI automatically handle dependencies hosted on GitHub by parsing a file called `dependencies.txt` placed in the project root directory. For correct parsing, GitHub dependencies' format must be of the forms provided in `dependencies.txt`:

```
com.github.Username:Repository:Branch-CommitHash
com.github.Username:Repository:ReleaseTag
```

where `CommitHash` may be replaced by `SNAPSHOT` to automatically select latest commits. For compilation through the Eclipse IDE, users should manually input dependencies in the `build.gradle` file. To distribute packages, users can create a Blang project with the CLI `create-blang-gradle-project`, and publish it in a GitHub repository.

11. Design patterns

This section discusses design patterns specific to programming in Blang. The goal of these design patterns is to enable users to design models going beyond Bayes nets, improve computational efficiency, and improve code readability.

11.1. Undirected graphical models

The mechanisms in Blang's default inference engine require the models to be in generative normal form. In some cases, in particular for users interested in undirected graphical models or Markov random fields (MRF), this may appear a stringent condition, since forward simulation in these models is computationally intractable.

We illustrate here a construction based on a type of “pseudo-prior”. Let $f(x) \propto \prod_{i \in I} \psi_i(x)$ denote an MRF, where I denotes a set of cliques that factorizes the MRF, and the subscripts i index factors of the respective cliques. We rewrite the model as $f(x) \propto f_0(x) \prod_{i \in I} \tilde{\psi}_i(x)$, where $f_0(x)$ is a “tractable” pseudo-prior. By tractable, we mean that we can sample and compute the normalization constant of the pseudo-prior. Annealing is then automatically performed on the factors $\prod_{i \in I} \tilde{\psi}_i(x)$ only, not on the pseudo-prior, ensuring finite marginalization for all interpolating distributions.

For example, consider the Ising model (Ising 1925) which is a type of MRF. In this case we use a product of independent Bernoulli random variables as a pseudo-prior. Note, we will make use of an “empty pipe symbol”, i.e., “`| IntVar first = ...`” which is explained after the example. We present the Blang code to implement the Ising model in Figure 24.

⁵⁴Under the hood, the mechanism for dependency management is **Maven**. However, GitHub repositories are seamlessly imported via **JitPack**. See <https://jitpack.io/> for details.

```

model Ising {
  param Double moment ? : 0.0
  param Double beta ? : log(1 + sqrt(2.0)) / 2.0
  param Integer N ? : 5
  random List<IntVar> vertices ? : latentIntList(N*N)

  laws {
    for (UnorderedPair<Integer, Integer> pair : squareIsingEdges(N)) {
      | IntVar first = vertices.get(pair.getFirst),
        IntVar second = vertices.get(pair.getSecond),
        beta
      ~ LogPotential({
        if ((first < 0 || first > 1 || second < 0 || second > 1))
          return NEGATIVE_INFINITY
        else
          return beta*(2*first-1)*(2*second-1)
      })
    }
    for (IntVar vertex : vertices) {
      vertex | moment ~ Bernoulli(logistic(-2.0*moment))
    }
  }
}

```

Figure 24: Ising model programmed in Blang, `Ising.bl`.

```

model LogPotential {
  param RealVar logPotential
  laws {
    logf(logPotential) {
      return logPotential
    }
  }
}

```

Figure 25: Implementation of log potential in Blang, `LogPotential.bl`.

Rather than using `logf` here for the likelihood, which would have violated the technical conditions for generative normal forms, we use the `LogPotential` utility in the SDK (shown for reference in `LogPotential.bl`).⁵⁵ Since `LogPotential` (Figure 25) does not define random variables, when it is invoked there are no variables to the left of the conditioning symbol in `| IntVar first = ...`. This follows naturally from the formal definition of composite laws. A second observation worthy of note is the conditioning of `| IntVar first = vertices.get(pair.getFirst)` as opposed to `| vertices`. This prevents the runtime architecture from assuming that these factors depend on the full `vertices` object, hence improving computational efficiency by a scaling proportional to the size of `vertices`. We emphasize this computational advantage in Section 11.2 with a Markov chain example.

11.2. Delayed graphical model construction

The runtime engine is able to decrease computational expense when it can detect sparsity patterns in models. This is handled automatically for simple objects but requires user input

⁵⁵https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/others/LogPotentialExample.bl

```

model MarkovChain {
  param Simplex initialDistribution
  param TransitionMatrix transitionProbabilities
  random List<IntVar> chain

  laws {
    chain.get(0) | initialDistribution ~ Categorical(initialDistribution)

    for (int step : 1 ..< chain.size) {
      chain.get(step) | IntVar previous = chain.get(step - 1),
        transitionProbabilities
        ~ Categorical({
          if (previous >= 0 && previous < transitionProbabilities.nRows)
            transitionProbabilities.row(previous)
          else
            invalidParameter
        })
    }
  }
}

```

Figure 26: Markov chains programmed in Blang, `MarkovChain.bl`.

for complex objects. For an example with a complex object `chain` consider the following Markov chain provided in Figure 26.⁵⁶

We condition on the previous step instead of the whole chain, using `chain.get(step) | IntVar previous = chain.get(step - 1)` as opposed to `chain.get(step) | chain`. This prevents the runtime architecture from computing factors involving the full `chain` object, potentially improving computational efficiency by a scaling proportional to the chain size. In other words, here the exact specification of the graphical model is delayed until the data is available.

In general, it is optimal to condition on the smallest possible scope. For example, suppose we have `SomeObject x` with conditional distribution on `ConditionalObject y`, where `y` has two `IntVar` fields `a` and `b`. If the distribution on `x` only requires the first field of `y`, `a`, then we should condition only on `a`. Hence, we use `x | IntVar v = y.a ~ Distribution(v)` as opposed to `x | y ~ Distribution(y.a)`. For a detailed understanding of this efficiency gain, we refer readers to Section 12.7.

11.3. Model reparameterization

It is often the case that distribution families can be written using different parameterizations, or that a family can be expressed as a special case of another family. Following “Don’t Repeat Yourself” (DRY) coding principles, the following pattern shows what is the best practice to express such reparameterizations.

To illustrate the pattern, consider how the Exponential distribution is coded in Figure 27 of the Blang SDK as a special case of the Gamma distribution.⁵⁷

⁵⁶https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/others/MarkovChainExample.bl

⁵⁷https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/others/ExponentialExample.bl

```

model Exponential {
  random RealVar realization
  param RealVar rate
  laws {
    realization | rate ~ Gamma(1.0, rate)
  }
}

```

Figure 27: Exponential distribution programmed in Blang, `Exponential.bl`.

```

model IntMixture {
  param Simplex proportions
  param List<IntDistribution> components
  random IntVar realization

  laws {
    logf(proportions, components, realization) {
      var sum = 0.0
      if (components.size != proportions.nEntries) {
        throw new RuntimeException
      }
      for (i : 0 ..< components.size) {
        val prop = proportions.get(i)
        if (prop < 0.0 || prop > 1.0) return NEGATIVE_INFINITY
        sum += prop * exp(components.get(i).logDensity(realization))
      }
      return log(sum)
    }
  }

  generate (rand) {
    val category = rand.categorical(proportions.vectorToArray)
    return components.get(category).sample(rand)
  }
}

```

Figure 28: Integer mixture model programmed in Blang, `IntMixture.bl`.

11.4. Distributions as parameters

In many situations, it is useful to have one or several parameters of a model to be themselves distributions. Consider for example a mixture model: it takes as input a list of distributions as well as mixture proportions, and creates a new distribution from it. Figure 28 provides an example of how this is implemented for mixtures of integer-valued distributions in Blang.

The `IntMixture` model can be invoked in the same manner as other models. An illustrative example using a mixture of two Poisson distributions is shown in Figure 29.⁵⁸

Here `Poisson::distribution(...)` is a convenient shortcut generated automatically: any model with only one random variable is automatically endowed with a `distribution(...)` function taking as input the model's parameters. The `distribution(...)` function returns a simplified application programming interface (API) for models having only one random variable. If that single random variable is of type `RealVar` (or `IntVar`), the returned value of `distribution(...)` is of type `RealDistribution` (or `IntDistribution`).⁵⁹ If the

⁵⁸https://github.com/UBC-Stat-ML/JSSBlangCode/blob/master/reproduction_material/example/jss/others/PoissonPoissonMixtureExample.bl

⁵⁹<https://www.stat.ubc.ca/~bouchard/blang/javadoc-dsl/blang/core/RealDistribution.html>,

```
x | lambda1, lambda2, pi
~ IntMixture(
  pi,
  #[Poisson::distribution(lambda1), Poisson::distribution(lambda2)]
)
```

Figure 29: Poisson mixture model coded in Blang, `PoissonPoissonMixtureExample.bl`.

type of the single random variable is neither `RealVar` nor `IntVar`, the returned value of `distribution(...)` is of the type `Distribution`.⁶⁰

12. Inference

Blang efficiently samples from posterior distributions by detecting sparsity patterns in the model, matching variable types with their associated roles in inference, then sample using state-of-the-art Monte Carlo methods.

In the following sections, we detail intermediate steps in the process described above. We first assume that a continuum of probability distributions is available. On one end of the spectrum, we have the posterior distribution, and the prior on the other. The prior is a distribution from which we can sample from assuming the model is in generative normal form. Then we describe the technical details used to automatically construct this continuum of interpolating probability distributions, along with invariant Markov chain kernels for each distribution in the interpolation.

12.1. Inference algorithms

Blang currently focuses on two complementary inference algorithms: sequential change of measure (SCM), and non-reversible parallel tempering (PT). SCM infers the exact posterior distribution asymptotically in memory, while PT infers the exact posterior distribution asymptotically in time. The former is an SMC algorithm and the latter a parallel MCMC algorithm.

A core concept present in both algorithms is the use of an adaptive sequence of tempered distributions extracted from a continuum interpolating from the prior to the posterior distribution. Through these tempering schemes, we are able to explore complex, multimodal distributions without the need for automatic differentiation; as such, these techniques are not limited to Euclidean spaces. For example, the default samplers for real and integer data types are their respective slice samplers (Neal 2003),⁶¹ which when used in a naive MCMC algorithm could perform poorly in highly correlated models. However in the context of SCM or PT, it is frequently the case that simple MCMC algorithms perform better than using specialized moves in a single chain (Ballnus, Hug, Hatz, Görlitz, Hasenauer, and Theis 2017). Furthermore, due to the inherent characteristics of these algorithms, they are trivially parallelized for efficient computing, and provide computation of model evidence at negligible cost.

<https://www.stat.ubc.ca/~bouchard/blang/javadoc-dsl/blang/core/IntDistribution.html>

⁶⁰<https://www.stat.ubc.ca/~bouchard/blang/javadoc-dsl/blang/core/Distribution.html>

⁶¹More precisely, a doubling and shrinking procedure is used as an adaptive scheme, whose details and validity are described and proved by Neal (2003).

For these reasons, SCM and PT are good candidates for automatic inference on generalized state spaces. These two algorithms can be used individually, but by default the SCM is used to initialize PT. This combination is motivated by the fact that SCM appears to often be better suited to quickly find a crude approximation. In particular SCM is able to find configurations of positive probability even in the presence of deterministic constraints (i.e., configurations having zero posterior probability). However, to obtain high quality samples, SCM may require a number of particles larger than what can be fitted in memory. PT on the other hand can provide approximations of arbitrary high quality without asymptotically infinite memory consumption.

Constructing a sequence of measures

Both SCM and PT inference algorithms require a continuum of measures. To retain theoretical guarantees, we must ensure each measure in this sequence has a finite normalization constant. To achieve this, we factorize our joint density into what we call likelihood $l_i(x)$ and prior $p_j(x)$ factors. Assuming a Blang model in generative normal form, the construction of such a continuum of probability measures begins with an exhaustive unrolling of composite laws to identify all atomic laws, or log factors. Each factor belongs to a model and as such each of its dependencies can be classified as either corresponding to **random** or **param**. If its dependency is **random**, we direct the corresponding edge in the factor graph as out-going. Otherwise, if it is a **param**, we direct the edge as in-coming. Likelihood factors are then defined as factors whose outgoing edges, if any, all connect to an observed variable; factors are classified as priors otherwise.

Suppose we have factorized our posterior as follows

$$\pi(x) \propto \prod_{i=1}^I l_i(x) \prod_{j=1}^J p_j(x),$$

where $l_i(x)$, $p_j(x)$ denote likelihood and prior factors respectively. As opposed to raising the product of likelihood and prior factors to some $t \in [0, 1]$, which may not yield a probability distribution, it is preferable to exponentiate the likelihood factors.

Additionally, it is common to have configurations of zero probability when performing inference over discrete combinatorial objects. In some scenarios, for example in pedigree analysis, these zero-valued likelihood evaluations can create difficulties in building irreducible samplers, thus invalidating convergence guarantees. We alleviate this restriction using the annealing scheme shown below,

$$\pi_t(x) = \frac{\gamma_t(x)}{Z_t} = \frac{\left(\prod_{i \in I} [l_i(x)]^t + \mathbb{I}(l_i(x) = 0) \epsilon_t \right) p(x)}{Z_t}, \quad (4)$$

where $\epsilon_t = \exp(-10^{100}t) \mathbb{I}(t < 1)$, $p(x) = \prod_{j=1}^J p_j(x)$, and $\mathbb{I}(\cdot)$ is the indicator function, i.e., $\mathbb{I}(\cdot)$ is equal to 1.0 when the argument is true, and 0.0 otherwise. We use the convention $0^0 = 0$ so that $\pi_0(x) = p(x)$. The conditions and effects of ϵ_t on the performance of algorithms have yet to be explored and is part of our future work. By design, the interpolating chains have a wide support (i.e., $p(x) > 0 \Rightarrow \pi_t(x) > 0$ for $t < 1$), while maintaining the guarantee of

having a finite normalization constant for all annealing parameters:

$$\begin{aligned}
\int \gamma_t(x) dx &= \int p(x) \prod_{i \in I} [(l_i(x))^t + \mathbb{I}(l_i(x) = 0)\epsilon_t] dx \\
&\leq \sum_{K:K \subset I} \epsilon_t^{|I|-|K|} \int p(x) \left(\prod_{i \in K} l_i(x) \right)^t dx \\
&= \sum_{K:K \subset I} \epsilon_t^{|I|-|K|} \int p(x) \left(\prod_{i \in K} l_i(x) \right)^t [\mathbb{I}(\prod_{i \in K} l_i(x) \geq 1) + \mathbb{I}(\prod_{i \in K} l_i(x) < 1)] dx \\
&\leq \sum_{K:K \subset I} \epsilon_t^{|I|-|K|} \left[\int p(x) \prod_{i \in K} l_i(x) dx + \int p(x) dx \right] \\
&< \infty
\end{aligned}$$

This proposed annealing scheme allows our sampler to traverse across multimodal distributions, preserve the correct marginal posterior distribution at room temperature (i.e., when $t = 1$), and guarantee convergence of normalizing constant estimates.

In a given execution of the PT and SCM inference algorithms, the full continuum of distributions $\{\pi_t : t \in [0, 1]\}$ is only instantiated on a finite grid $0 = t_0 < t_1 < \dots < t_N = 1$, called an *annealing schedule*. Since the performance of both PT and SCM are sensitive to the choice of annealing schedules, they each use a specialized algorithm to automatically optimize the annealing schedule (described in the next sections). To perform a continuous optimization over $(t_1, t_2, \dots, t_{N-1})$ with monotonicity constraints, the algorithms rely on the fact that the discrete sequence of distributions is embedded in a continuum of distributions.

Sequential change of measure

Informally, **Blang**'s SCM inference engine initializes a population of particles from a prior distribution, and iteratively perturbs and reweights the particles. The number of particles used is 1000 by default and can be set using the `--engine.nParticles` option. More precisely, SCM is a special case of the sequential Monte Carlo (SMC) sampler (Del Moral *et al.* 2006, Section 3.3.2.3) combined with an adaptive tempering schedule described by Zhou *et al.* (2016) (also called ‘‘annealed SMC’’, e.g., in Wang, Wang, and Bouchard-Côté 2020). SMC *samplers* are an extension or generalization of SMC methods, which allow for sampling from a sequence of distributions defined on a fixed state space, as opposed to the more common SMC setup (Doucet and Johansen 2011) consisting of product spaces of increasing dimensionality.

As described in Del Moral *et al.* (2006, Section 3.3.2.3) the proposals we use consist in MCMC kernels targeting each of the intermediate distributions (see Section 12.7 for a detailed description on their construction). Del Moral *et al.* (2006) also justify in this context the incremental weight updates given by

$$w_t(x_{t-1}, x_t) = \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})}, \quad (5)$$

where $\gamma_t(x)$ is the numerator in the right-hand side of Equation 4. Initialization is done using the prior sampler described in Section 12.6.

Due to the weight degeneracy problem reviewed in Doucet and Johansen (2011), a resampling procedure is required. Resampling prevents the population of sample weights from

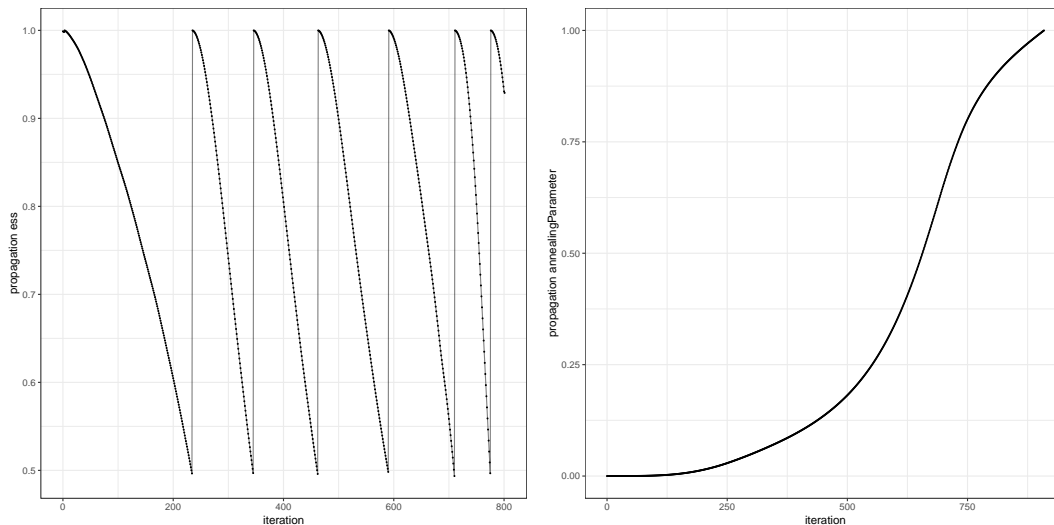


Figure 30: SCM monitoring plots automatically created when SCM is used and the default post-processing tool activated (`--engine SCM --postProcessor DefaultPostProcessor`). Here we show examples for the Gaussian mixture model in Section 6, using 10000 particles. The adaptive resampling scheme based on ESS estimates is performed by default for SCM. Left: ESS plotted against iterations, automatically created in `monitoringPlots/propagation-ess.pdf`. Each “spike” observed in this plot correspond to a resampling step taking place. When ESS falls beneath a predefined threshold, 0.5 here, particles are resampled to prevent weight degeneracy. This resampling procedure “refreshes” the relative ESS to 1.0. Right: The resulting adaptive annealing schedule, found in `monitoringPlots/propagation.pdf`. The y -axis corresponds to annealing parameters, and the x -axis to the iteration number.

collapsing into a point mass. However, resampling injects additional noise into the sampling process. Motivated by the need to balance these two factors, we use a standard procedure to adaptively determine when resampling should be performed. The effective sample size (ESS) is computed at each iteration (Kong 1992). If the relative ESS – ESS divided by population size – falls beneath a predetermined threshold, resampling is performed. Figure 30 (left) illustrates the resampling procedure’s effect on ESS. This threshold value defaults to 0.5 in Blang, and can be set, for example, to 0.4 using the command-line argument `--engine.resamplingESSThreshold 0.4`. By default, the resampling scheme used is the stratified sampling of Kitagawa (1996) (use `--engine.resamplingScheme MULTINOMIAL` for multinomial resampling).

SMC samplers rely on a discrete set of interpolating distributions $0 = t_1 < t_2 < \dots < t_N = 1$. As initially proposed in Jasra, Stephens, Doucet, and Tsagaris (2011) and improved in Zhou *et al.* (2016), instead of building this sequence a priori, we construct it incrementally and adaptively. At each step the next annealing parameter is determined so as to cause a fixed decay in the relative conditional ESS as defined in Zhou *et al.* (2016). Figure 30 (right) shows an example of a resulting adaptive annealing schedule. Finding the next annealing parameter is a simple univariate root finding problem. Since the weight update shown in Equation 5 does not depend on x_t , only the already available particles from the previous iteration, x_{t-1} , the com-

putational cost of the root finding problem is negligible. By default, the targeted decay is set to 0.9999 and can be controlled via `--engine.temperatureSchedule.threshold`. Setting it to a lower value will speed up computation at the cost of a less accurate posterior distribution (and vice versa). One disadvantage of this adaptation scheme is that the running time of the method is random and may be hard to predict a priori. If the user requires a prespecified number of iterations, adaptive construction of the sequence of distribution can be turned off, for example to use a fixed number of 20 iterations and an equally spaced annealing schedule $0 = t_1 < t_2 < \dots < t_{20} = 1$, use `--engine.temperatureSchedule FixedTemperatureSchedule --engine.temperatureSchedule.nTemperatures 20`. Custom mechanisms to control the schedule can be added by implementing the interface `TemperatureSchedule`.⁶² If for example the user implements their own algorithm in a class called `MySchedule` located in package `mypackage`, to enable its use during inference, add the command line arguments `--engine.temperatureSchedule mypackage.MySchedule`.

After SCM inference is performed, `Blang` performs one last round of resampling followed by 5 rounds of particle rejuvenation on each particle. This results in a set of equally weighted particles. The amount of rejuvenation to perform after the final resampling round can be controlled via `--engine.nFinalRejuvenations`, for example use `--engine.nFinalRejuvenations 10`, for 10 rejuvenation rounds.

Non-reversible parallel tempering

`Blang` incorporates a non-reversible, adaptive parallel tempering (PT) algorithm (Syed *et al.* 2019). PT (Geyer 1991) is an MCMC method that operates on product spaces. Informally, PT runs N Markov chains in parallel, each targeting a distribution from a sequence of tempered (i.e., annealed) distributions indexed by $0 \leq t \leq 1$ (used when constructing a sequence of measures). Each PT iteration consists of two phases: a local exploration phase taking place within individual chains, and a communication phase taking place between chains.

In the local exploration step, for chains with $t > 0$, the state is updated via samplers or MCMC kernels invariant with respect to the chain's target distribution (the construction of these kernels is detailed in Section 12.7). For the chain with $t = 0$, the local exploration step consists in an independent draw from the prior distribution (the construction of the independent sampler for the prior is detailed in Section 12.6). If the user requires using MCMC samplers for $t = 0$ instead of prior sampling, the option `--engine.usePriorSamples false` can be used.

In the communication phase, swaps between neighbor chains are proposed and accepted/rejected according to the Metropolis-Hastings ratio. Informally, this swapping procedure provides opportunities for states to traverse across modes, as the prior allows independent sampling and hence a form of regeneration. Even when sampling from the prior is not possible, annealing often yields MCMC kernels with better mixing rates.

In our implementation, both the exploration and communication phases are parallelized in the number of parallel chains N (see Section 12.3 for details). However to leverage this parallelism, following the theoretical analysis of Syed *et al.* (2019), special attention has been devoted (1) to the details of how the swap mechanism is performed, and (2) to the tuning of the annealing schedule t_1, t_2, \dots, t_{N-1} introduced in the last section.

⁶²<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/engines/internals/schedules/TemperatureSchedule.html>

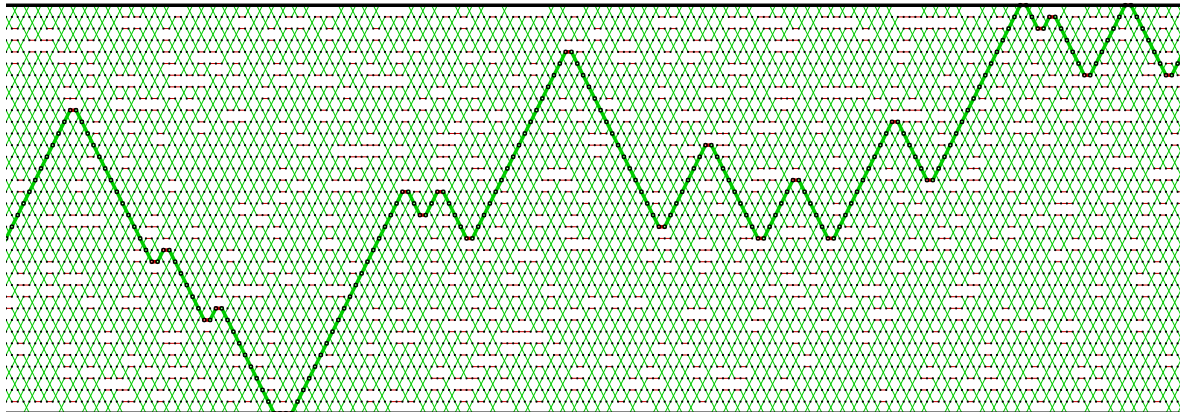


Figure 31: Visualization (cropped) of the chain swaps proposed while running non-reversible PT (add `--postProcessor.runPxviz true --postProcessor.boldTrajectory 1` to create this visualization). The x -axis corresponds to PT iterations, and the y -axis corresponds to different parallel chains, with the one at the top corresponding to the posterior distribution, the one at the bottom, to the prior distribution, and those in between interpolating between the two. When a swap is accepted (green line segments), two chains exchange their states, denoted by crossing lines. When a swap is rejected we use red line segments. An *index process* is obtained by considering a path formed by these line segments (one index process is shown as a bold line for ease of interpretation). An *annealed restart* is defined as a path segment within an index process which starts at the prior and ends at the posterior.

Point (1) is motivated by a sharp contrast between the performance of reversible and non-reversible flavors of PT. Performance in the following discussion is based on the notion of *annealed restarts*, defined along with the related notion of the index process in Figure 31. We define PT performance as the fraction of iterations where an annealed restart is just completed at the current iteration. This is called the restart rate, which we denote by τ , and it is equivalent (up to an additive factor of 1) to the notion of round trip rate popular in the PT literature (Katzgraber, Trebst, Huse, and Troyer 2006; Lingenhil, Denschlag, Mathias, and Tavan 2009).

Previous theoretical work has focused on reversible PT where the groups of chains to swap are selected at random. In the reversible regime, several lines of work (Rathore, Chopra, and de Pablo 2005; Atchadé, Roberts, and Rosenthal 2011) have demonstrated that even when a high number of cores is available, one still has to ensure that N , and hence the number of cores leveraged, is not too large. More precisely, the performance of reversible PT collapses as N increases, even when communication and local exploration are fully parallel. For example the results in Atchadé *et al.* (2011) imply that $\tau_{\text{rev},N} \rightarrow 0$ as $N \rightarrow \infty$. Surprisingly, this performance collapse disappears when a non-reversible flavor of PT is used: Syed *et al.* (2019) identified conditions where $\tau_{\text{non-rev},N} \rightarrow c$ as $N \rightarrow \infty$; the constant $c > 0$ (model-dependent) is discussed further below. Even more surprising is that algorithmically, the distinction needed to make PT non-reversible is minimal: it is simply the use of a deterministic alternation of two specific types of swap kernels, those swapping $i, i + 1$ with i even, followed by similar swaps with i odd. This algorithm can be traced back to Okabe, Kawata, Okamoto, and Mikami (2001), however, only recently its non-reversible dynamics have been identified and

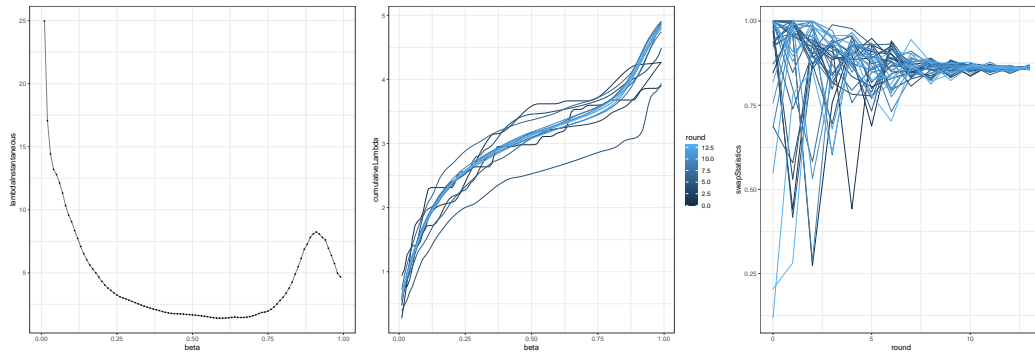


Figure 32: Left: final estimate of the local communication barrier $\hat{\lambda}$ versus the annealing parameter t (labeled “beta” in the Blang output). The spiking phenomenon around $t = 0.9$ is indicative of a phase transition. This corresponds to the mixture indicator variables going from a disorganized configuration (the side of the peak at $t = 0.9$ closer to the prior on the left) to a clustered configuration (the side of the peak closer to the posterior). Middle: estimates of the cumulative communication barrier, $\hat{\Lambda}(t)$, with each color corresponding to a different iterative round of the annealing schedule optimization algorithm. Right: here each line (color) is one of the N chains, and the line tracks the average acceptance probability (ordinate) between that chain and its neighbor for each round of the schedule optimization algorithm (abscissa). In contrast to reversible PT, NRPT does not need to restrict swap acceptance probability to low values such as the 23% acceptance rule of [Atchadé et al. \(2011\)](#).

used to prove the existence of a qualitative gap between the reversible and non-reversible flavors of PT. The gap can be established both non-asymptotically ($\tau_{\text{rev},N} < \tau_{\text{non-rev},N}$ for all N), and also asymptotically as $N \rightarrow \infty$, in which case the performance of non-reversible PT, $\tau_{\text{non-rev},N}$, is furthermore guaranteed to be monotonically increasing for N large enough.

More importantly, non-reversibility opens the door for highly parallel algorithms to optimize over the annealing schedule, hence addressing point (2) above. By default, Blang’s PT engine uses the non-reversible schedule optimization from ([Syed et al. 2019](#), labeled NRPT henceforth). In contrast, at the time of writing, mainstream probabilistic programming languages either lack support for parallel tempering ([Plummer 2003](#); [Lunn et al. 2012](#); [Salvatier et al. 2016](#); [Carpenter et al. 2017](#)), or require manual input of the annealing parameters ([Foreman-Mackey, Hogg, Lang, and Goodman 2013](#)).

To outline how the NRPT algorithm works, we first outline the asymptotic distribution of a single index process (for example, the bold line in [Figure 31](#)) as $N \rightarrow \infty$. While for reversible PT this distribution converges to a diffusion, for non-reversible PT, it converges to a piecewise-deterministic Markov process (PDMP). See [Davis \(1993\)](#) for background on PDMPs. The rate parameter of this limiting PDMP, λ , a positive function taking as input an annealing parameter $t \in [0, 1]$, can be interpreted as being proportional to the expected rejection rate for a swap between π_t and $\pi_{t+\epsilon}$. See [Figure 32](#) (left) for an example of an estimate $\hat{\lambda}$ from the model in [Section 6](#). Moreover, the constant c introduced earlier as the asymptotic non-reversible performance, $\tau_{\text{non-rev},N} \rightarrow c$ can be written as $c = (2 + \Lambda)^{-1}$, where $\Lambda(t) = \int_0^t \lambda(t') dt'$. We call λ , $\Lambda(t)$, and $\Lambda = \Lambda(1)$ the local, cumulative and global communication barriers respectively.

Importantly, all three communication barriers can be estimated from the MCMC output, and

used as the basis for tuning PT as described in Syed *et al.* (2019). First, Λ can be used as a measure of the difficulty of PT-based inference for a given model: this is supported by its relation with the model-specific constant c described earlier. We recommend to use a number of chains proportional to Λ . Using at least 2Λ appears to provide a good starting point empirically. Second, NRPT uses $\Lambda(t)$ to optimize the annealing parameters using the following strategy. The algorithm iteratively estimates $\hat{\Lambda}(t)$, using a simple and asymptotically consistent rule $\hat{\Lambda}(t_i) = \sum_{j=1}^i \hat{r}^{(j-1,j)}$, and $\hat{\Lambda}(\cdot)$ interpolated using a monotone cubic spline between the t_i 's, where $\{t_i\}$ are annealing parameters from the previous iterations, and $\hat{r}^{(j-1,j)}$ is the empirical swap rejection rate, also obtained from the previous iteration. The algorithm then computes univariate quantile of $t \mapsto \hat{\Lambda}(t)/\hat{\Lambda}(1)$ to update the annealing schedule. This is repeated using a doubling scheme, where the first round uses 1 iteration to estimate $\hat{\Lambda}(t)$, followed by annealing parameters update, the second round uses 2 iterations based on the updated schedule, followed by an update of the annealing parameters, then 4, 8, etc. See Figure 32 (middle) for an example of how estimates of $\hat{\Lambda}$ progress as the number of rounds increases. As a byproduct of the NRPT algorithm we obtain a burn-in mechanism: by default, all post-processing uses only the samples produced by the last optimization round which is equivalent to a 50% burn-in. The only exception is for the trace plot, which is shown for both the whole MCMC trace in the output folder `tracePlotsFull`, and for the post burn-in phase, `tracePlots`. The default of 50% burn-in can be customized via `--postProcessor.burnInFraction`.

Alternative mechanisms can be used to control the annealing parameters. By default, the initial annealing schedule is uniform, other initial values can be used, see `--engine PT --help` for various options. Optimization of the annealing parameters can be disabled with the argument `--engine.adaptFraction 0.0`. Custom mechanisms to control the initial schedule can be added by the user by implementing the interface `TemperatureLadder`.⁶³ If for example the user implements their own algorithm in a class called `MyLadder` located in package `mypackage`, to enable it, use `--engine.ladder mypackage.MyLadder`.

One tuning parameter that can be used to speed-up the execution of NRPT is the expected number of times each local exploration kernel should be used between two rounds of swap attempts, `--engine.nPassesPerScan` (fractional values are accepted). By default, this is set to 3, so that a theoretical assumption called effective local exploration (ELE, Syed *et al.* 2019), is well approximated. However, we observed that performance was robust to this choice so if the local exploration kernels are reasonably efficient, lower values will lead to similar behavior of the index processes for a lower computational budget. Conversely, if the local exploration kernels perform very poorly, it may be useful to explore higher values for the argument `--engine.nPassesPerScan`.

If the hard-drive space required to store the samples produced by PT becomes prohibitive, one option is to enable thinning by providing an input `--engine.thinning` greater than one. For example, `--engine.thinning 2` will store samples only once every two PT iterations. An alternative (available for all engines), is to compress the samples in `.gz` format, which is enabled using `--experimentConfigs.tabularWriter.compressed true`. All post-processing is compatible with the compressed samples format.

Initialization of PT is by default performed by first running SCM using an annealing schedule

⁶³<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/engines/internals/ladders/TemperatureLadder.html>

containing all annealing parameters in PT's initial schedule. The SCM initialization can be configured using the same arguments as those described when discussing the sequential change of measure but with the prefix `engine.scmInit`. For example, to increase the number of particles (set to 100 for the initialization run), use `--engine.scmInit.nParticles 200`.

Other inference engines

When a model is not in generative normal form, the PT and SCM engines cannot be used. In such case, the user can still run a basic, single chain MCMC via `--engine MCMC`. This option is essentially a shortcut for setting the PT engine to use a single chain, to avoid using SCM for initialization, and to avoid other checks that assume a generative normal form. The PT command line arguments from non-reversible parallel tempering that are relevant to single-chain MCMC can still be used, in particular `--engine.nScans` and `--engine.thinning`.

Another convenient shortcut is `--engine AIS` which uses SCM but with resampling disabled. This is known as the annealed importance sampling algorithm (AIS, Neal 2001). The SCM arguments relevant to AIS can still be used with this engine, namely `--engine.nParticles`, `--engine.nFinalRejuvenation`, and `--engine.temperatureSchedule.threshold`.

In cases where the user would like to sample independent and identically distributed realizations from a model where no observation is present, the engine `--engine Forward` (for forward sampling) with option `--engine.nSamples 1` can be used.

When all random variables in a small model are discrete, the argument `--engine Exact` will enumerate all possible scenarios. Note that the `DefaultPostprocessor` should not be used to analyze the output of the exact engine. This is because the output in the folder `samples` has a different interpretation than with the other engines: instead of representing equally weighted samples, they represent weighted samples with weight indicated in a row called `logProbability`.

Finally, the inference engine can be customized. This is achieved by implementing the interface `PosteriorInferenceEngine`.⁶⁴ If for example the user implements their own algorithm in a class called `MyEngine` located in package `mypackage`, to enable its use during inference, add the command line arguments `--engine mypackage.MyEngine`.

12.2. Pseudo-random generator

The current pseudo-random generator is the Mersenne Twister generator (Matsumoto and Nishimura 1998) as implemented in the `MathCommons` package. By default, the seed 1 is used. For inference engines based on randomized algorithms (all current algorithms except `Exact`), this can be changed using the command line argument `--engine.random` followed by an integer.

12.3. Parallelization

Due to the nature of PT and SCM algorithms, parallelization can be used to obtain significant performance improvements. In both PT and SCM, transition MCMC kernels are applied in parallel across particles/chains. In addition to parallelization of transition kernels, PT also performs its swap operations in parallel.

⁶⁴<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/engines/internals/PosteriorInferenceEngine.html>

Blang uses lightweight threads to parallelize these operations (Friesen 2015). Specifically, it uses the algorithm described in Leiserson, Schardl, and Sukha (2012) as implemented in Steele and Lea (2013). This implementation allows each chain to pertain to its own random stream, consequently avoiding any blocking between threads. Furthermore, this implementation implies any numerical output will not be altered by the number of threads utilized given fixed random seeds.

For controlling multi-threading, use `--engine.nThreads Max` to take advantage of as many threads as there are cores in the host machine, `--engine.nThreads Dynamic` to dynamically allocate threads based on the overall system usage (the default behavior, which ensures analysts can smoothly carry out other tasks while inference is running in the background), `--engine.nThreads Single` to force single-thread mode, and `--engine.nThreads Fixed` `--engine.nThreads.number 2` to fix a specific number of threads to use.

12.4. Marginal likelihood computation

Standard Bayesian model selection requires computing the marginal likelihood, also known as the evidence. The marginal likelihood is conceptually simple: it is the probability or the density of the observed data. However computing or approximating this single scalar is often challenging. Fortunately, both PT and SCM automatically compute the marginal likelihood with no extra computational cost.

Our PT engine supports estimation of the marginal likelihood through two methods: thermodynamic integration (Ogata 1989), and the stepping stone estimator (Xie, Lewis, Fan, Kuo, and Chen 2011). For models with hard constraints (i.e., models whose likelihood is equal to zero for particular configurations of states proposed by the sampling algorithm), the technical conditions underlying thermodynamic integration may not be satisfied, and that estimator is automatically omitted in such cases. The stepping stone estimator can still be used in these cases. In SCM, the evidence comes as a by-product of the weights computed by the algorithm, see, e.g., Del Moral *et al.* (2006).

In contrast to Blang, other mainstream probabilistic programming languages require additional packages and external dependencies to approximate the marginal likelihood. For example in Stan, one would require additional post-processing with bridge sampling (Meng and Wong 1996) using packages such as **bridgesampling** (Gronau and Singmann 2021; Gronau, Singmann, and Wagenmakers 2020).

12.5. Diagnostics

We summarize here some diagnostic strategies that can be used to assess the quality of the posterior distribution approximation. With the PT inference engine, the key diagnostics are the ESS estimates (Section 4) and the number of annealed restarts (see Figure 31 for the definition, and `monitoring/actualTemperedRestarts.csv` for the estimates). Each annealed restart incorporates a unique independent draw from the prior chain successfully propagated to the posterior chain. This can be complemented with inspection of the trace plots (Section 4). Finally, another strategy is to monitor the marginal likelihood: a separate estimate is provided for each adaptation round in the PT engine (in `monitoring/logNormalizationContantProgress.csv`), so its convergence can be readily monitored, and moreover one can check the agreement of PT's estimate with the orthogonal marginal likelihood estimator used by SCM (either based on the automatic SCM initialization, or from a separate run; see `logNormalizationEstimate.csv`).

12.6. Construction of prior samplers

Consider the sequence of distribution in Equation 4 at $t = 0$, where we recover the prior distribution $p(x)$. When a model is in generative normal form (Section 5.7), **Blang** automatically constructs an efficient algorithm to sample from the prior distribution $p(x)$. Briefly, the normal form property guarantees that we can orient the factor graph over the latent variables into a directed graphical model. The generative normal form property enables the enumeration of forward samplers provided by the `generate` blocks. Finally, as a preprocessing step, we order these `generate` blocks according to a linearization of the directed graphical model.

12.7. Construction of invariant samplers

We first describe how a **Blang** model m is transformed into an efficient representation aware of m 's sparsity patterns. The transformed representation is an instance of `SampledModel`,⁶⁵ a mutable object keeping track of the state space and offering methods to: (1) change the annealing parameter of the model, (2) apply a transition kernel in place targeting the current annealing parameter, (3) perform forward simulation in place, (4) obtain the joint log density of the current configuration, and (5) duplicate the state via a deep cloning library (Appendix C).

Preprocessing

The process of translating m into a `SampledModel` begins with the instantiation of model variables. After this is done, a list l of factors is recursively constructed. That is, we recursively search through m for sub-models, and terminate when we have identified and added all atomic laws to l .

The next phase of initialization consists of building an *accessibility graph* between all objects in a model,⁶⁶ defined as follows: the set of vertices is the set of objects defined by a model, starting at the root model, and of the constituents of these objects recursively. Constituents are fields in the case of objects and integer indices in the case of arrays. Constituents can also be customized, for example, in order to index entries of matrices. The directed edges of the accessibility graph connect objects to their constituents, and constituents to the object they resolve to, if any. We say that object o_2 is accessible from o_1 if there is a directed path from o_1 to o_2 in the accessibility graph.

Once the accessibility graph has been constructed, the latent variables in m are extracted from the vertex set of the accessibility graph. These variables are the intersection of objects of a type annotated with `@Samplers` and objects that are mutable, or have accessible mutable children. Mutability here corresponds to the class having fields that are either arrays or that are *non-final* (the latter being the Java terminology, or equivalently, in **Xtend**, fields not marked by `val`). In other words, latent variables are objects that have a designated sampler, and are or have access to non-final fields. Immutability is therefore the main mechanism used to define observed (fixed) values. Additionally, we can mark indices in matrices and arrays as observed. This is accomplished by the `Observations` object.⁶⁷ For example, `observationsObject.markAsObserved(mtx.getRealVar(i, j))` marks entry (i, j) of

⁶⁵<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/runtime/SampledModel.html>

⁶⁶Here objects refer to the same objects as defined in Java, i.e., dynamically allocated class instances or arrays.

⁶⁷<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/runtime/Observations.html>

matrix `mtx` as observed. In scenarios where objects or fields are accessible but unused in factors, the exploration of such objects and fields can be skipped. This can be handled in the construction of the accessibility graph by using the annotation `@SkipDependency`. With our accessibility graph constructed and latent variables identified, we can now exploit the model's sparsity patterns by constructing a factor graph.

Exploiting sparsity

Samplers can be made more efficient by avoiding unnecessary computation of model components; we exploit a model's sparsity pattern by building factor graphs via linear time graph algorithms on our accessibility graph.

Given a latent variable v and factor f , we can determine whether the application of a sampling operator on v can change the numerical value of the factor f . This is accomplished by assessing v and f 's co-accessibility. Two objects o_1 and o_2 are said to be co-accessible if there is a mutable object (as defined previously) o_3 such that o_3 is accessible from both o_1 and o_2 . Through this awareness of sparsity patterns, we can now perform sampling operations on variables without computing every factor involved in the model. The cost of the entire preprocessing procedure has negligible cost in comparison with the performance to be gained from its implications.

For a concrete example, we refer readers to the Markov chain example (Section 11.2).

Matching transition kernels (samplers)

Once sparsity patterns have been identified, samplers are matched to latent variables through the `@Samplers` annotation. We have seen examples of this annotation in the permutation example of Section 9. Here we provide more details on this process.

Given a certain model we would like to sample from, the first step is to identify which types of samplers are needed. To do so, recall that each latent variable is by definition of a type annotated with the `@Samplers` annotation. Now the `@Samplers` annotation is required to include as arguments a list of types responsible for sampling that type. For example, the class `DenseSimplex` is annotated with `@Samplers(SimplexSampler)`.⁶⁸ Repeating this process for each type of latent variable, this gives us a pool of sampler types. This pool of sampler types is summarized at the top of the standard output when sampling is performed, for example,

2 samplers constructed with following prototypes:

```
RealScalar sampled via: [RealSliceSampler]
IntScalar sampled via: [IntSliceSampler]
```

The next step is to attempt to instantiate one `Sampler` object for each latent variable. We will walk through this process based on the example of instantiating a `SimplexSampler`, shown in Figure 33.⁶⁹

⁶⁸It is also possible to add or exclude samplers from the command line, using the options `--samplers.additional` and `--samplers.excluded` respectively, followed by fully qualified Java types pointing to sampler implementations. If the user wants to only rely on the set of samplers specified in the command line and not the ones obtained from the annotation, the option `--useAnnotation false` can be used.

⁶⁹<https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/mcmc/SimplexSampler.xtend>

```

...
class SimplexSampler implements Sampler {
    @SampledVariable DenseSimplex simplex
    @ConnectedFactor List<LogScaleFactor> numericFactors
    @ConnectedFactor Constrained constrained

    override void execute(Random rand) { ... }
    override boolean setup(SamplerBuilderContext context) { ... }
}

```

Figure 33: Snippet of the simplex sampler, `SimplexSampler.xtend`.

At this point of the process, we have on the one hand a specific instance of a simplex random variable s within a factor graph, and on the other hand, the `SimplexSampler` type. We first look at all the factors connected to s that can be assigned to fields annotated by `@ConnectedFactor` in `SimplexSampler`. For example, if any number of factors of type `LogScaleFactor` are encountered, it can be assigned by adding it to the list `numericFactors`. Similarly, if s is connected to no more than one `Constrained` factor, that factor can be assigned to the field `constrained`. If all neighbors of s can be matched in this fashion, then all these annotated fields will be automatically populated by the neighbor factors of s . If this is not the case, i.e., if some neighbor factor cannot be assigned to one of the fields marked by `@ConnectedFactor`, then the current sampler type will not be matched with s .

Provided we have a match, then the field `simplex` annotated with `@SampledVariable` is automatically initialized with the sampled variable. After this is done, the `setup` method is invoked to (1) perform any required pre-computation, and (2) to provide a chance to reject matching the sampler based on more complex criteria (for example, if deciding whether it is possible for this sampler to handle sampling s based on information only available at runtime; the return value of `setup` determines if the sampler will be instantiated in the given context). The function `setup` is provided as input an object of type `SamplerBuilderContext` which makes it possible to use more fine grain information on the factor graph.⁷⁰ For example, if s is itself composed of several objects s_1 and s_2 to be sampled one after the other, it would be possible to cache a more precise decomposition of the list of factors connected to each by using `context.connectedFactors(new ObjectNode(s1))`.

This completes the setup phase. Next, during Monte Carlo sampling, the variable being sampled is updated in place through the invocation of `execute`.

Going back to the simplex example, the `Constrained` factor is used here to indicate that the sampler being constructed is aware of the constraints posed by simplex variables.

Computational details

All the programs in this paper were run using `blangSDK` 2.13.1 on a Mac OS X version 10.14.5. The device used is a MacBook Pro (15-inch, 2018) with a 2.2 GHz 6-core Intel Core i7 processor and a 2.2 GHz Radeon Pro 555X graphics card and 32 GB of 2400 MHz DDR4 memory.

⁷⁰<https://www.stat.ubc.ca/~bouchard/blang/javadoc-sdk/blang/mcmc/internals/SamplerBuilderContext.html>

Acknowledgments

This research was supported by a Discovery Grant from the National Science and Engineering Research Council, a Canadian Statistical Sciences Institute Collaborative Research Team Project, and Michael Smith Foundation RRF Grant.

References

- Ackerman NL, Freer CE, Roy DM (2017). “On Computability and Disintegration.” *Mathematical Structures in Computer Science*, **27**(8), 1287–1314. doi:10.1017/s0960129516000098.
- Atchadé YF, Roberts GO, Rosenthal JS (2011). “Towards Optimal Scaling of Metropolis-Coupled Markov Chain Monte Carlo.” *Statistics and Computing*, **21**(4), 555–568. doi:10.1007/s11222-010-9192-1.
- Ballnus B, Hug S, Hatz K, Görlitz L, Hasenauer J, Theis FJ (2017). “Comprehensive Benchmarking of Markov Chain Monte Carlo Methods for Dynamical Systems.” *BMC Systems Biology*, **11**, 63. doi:10.1186/s12918-017-0433-1.
- Bezanson J, Edelman A, Karpinski S, Shah VB (2017). “Julia: A Fresh Approach to Numerical Computing.” *SIAM Review*, **59**(1), 65–98. doi:10.1137/141000671.
- Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman ND (2018). “Pyro: Deep Universal Probabilistic Programming.” *Journal of Machine Learning Research*, **20**(1), 973–978.
- Bürkner PC, Vuorre M (2019). “Ordinal Regression Models in Psychology: A Tutorial.” *Advances in Methods and Practices in Psychological Science*, **2**(1), 77–101. doi:10.1177/2515245918823199.
- Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017). “Stan: A Probabilistic Programming Language.” *Journal of Statistical Software*, **76**(1), 1–32. doi:10.18637/jss.v076.i01.
- Carter B, McCrea WH (1983). “The Anthropic Principle and Its Implications for Biological Evolution.” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **310**(1512), 347–363. doi:10.1098/rsta.1983.0096.
- Chen MH, Shao QM (1999). “Monte Carlo Estimation of Bayesian Credible and HPD Intervals.” *Journal of Computational and Graphical Statistics*, **8**(1), 69–92. doi:10.1080/10618600.1999.10474802.
- Davis MHA (1993). *Markov Models & Optimization*. Chapman and Hall.
- Del Moral P, Doucet A, Jasra A (2006). “Sequential Monte Carlo Samplers.” *Journal of the Royal Statistical Society B*, **68**(3), 411–436. doi:10.1111/j.1467-9868.2006.00553.x.
- Doucet A, Johansen AM (2011). “A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later.” In D Crisan, B Rozovskii (eds.), *The Oxford Handbook of Nonlinear Filtering*, pp. 656–704. Oxford University Press.

- Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987). “Hybrid Monte Carlo.” *Physics Letters B*, **195**(2), 216–222. doi:10.1016/0370-2693(87)91197-x.
- Efftinge S, Eysholdt M, Köhnlein J, Zarnekow S, von Massow R, Hasselbring W, Hanus M (2013). “**Xbase**: Implementing Domain-Specific Languages for **Java**.” *ACM SIGPLAN Notices*, **48**(3), 112–121. doi:10.1145/2480361.2371419.
- Efftinge S, Völter M (2006). “oAW xText: A Framework for Textual DSLs.” *Proceedings of Workshop on Modeling Symposium at Eclipse Summit*.
- Flegal JM, Jones GL (2010). “Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo.” *The Annals of Statistics*, **38**(2), 1034–1070. doi:10.1214/09-aos735.
- Foreman-Mackey D, Hogg DW, Lang D, Goodman J (2013). “**emcee**: The MCMC Hammer.” *Publications of the Astronomical Society of the Pacific*, **125**(925), 306–312. doi:10.1086/670067.
- Friesen J (2015). *Java Threads and the Concurrency Utilities*. Apress, Berkeley.
- Geweke J (2004). “Getting It Right: Joint Distribution Tests of Posterior Simulators.” *Journal of the American Statistical Association*, **99**(467), 799–804. doi:10.1198/016214504000001132.
- Geyer CJ (1991). “Markov Chain Monte Carlo Maximum Likelihood.” *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, Interface Foundation, Fairfax Station*, pp. 156–163.
- Goodman ND, Mansinghka VK, Roy DM, Bonawitz K, Tenenbaum JB (2008). “**Church**: A Language for Generative Models.” In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI’08*, pp. 220–229.
- Greiner J, Burgess JM, Savchenko V, Yu HF (2016). “On the Fermi-GBM Event 0.4 s after GW150914.” *The Astrophysical Journal*, **827**(2), L38. doi:10.3847/2041-8205/827/2/L38.
- Griffiths TL, Ghahramani Z (2011). “The Indian Buffet Process: An Introduction and Review.” *Journal of Machine Learning Research*, **12**(32), 1185–1224.
- Gronau QF, Singmann H (2021). **bridgesampling**: *Bridge Sampling for Marginal Likelihoods and Bayes Factors*. R package version 1.1-2, URL <https://CRAN.R-project.org/package=bridgesampling>.
- Gronau QF, Singmann H, Wagenmakers EJ (2020). “**bridgesampling**: An R Package for Estimating Normalizing Constants.” *Journal of Statistical Software*, **92**(10), 1–29. doi:10.18637/jss.v092.i10.
- Höhna S, Heath TA, Boussau B, Landis MJ, Ronquist F, Huelsenbeck JP (2014). “Probabilistic Graphical Model Representation in Phylogenetics.” *Systematic Biology*, **63**(5), 753–771. doi:10.1093/sysbio/syu039.

- Höhna S, Landis M, Heath T, Boussau B, Lartillot N, Moore B, Huelsenbeck J, Ronquist F (2016). “**RevBayes**: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.” *Systematic Biology*, **65**(4), 726–736. doi:10.1093/sysbio/syw021.
- Ising E (1925). “Beitrag zur Theorie des Ferromagnetismus.” *Zeitschrift für Physik*, **31**(1), 253–258. doi:10.1007/bf02980577.
- Jasra A, Holmes CC, Stephens DA (2005). “Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling.” *Statistical Science*, **20**(1), 50–67. doi:10.1214/088342305000000016.
- Jasra A, Stephens DA, Doucet A, Tsagaris T (2011). “Inference for Lévy-Driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo.” *Scandinavian Journal of Statistics*, **38**(1), 1–22. doi:10.1111/j.1467-9469.2010.00723.x.
- Katzgraber HG, Trebst S, Huse DA, Troyer M (2006). “Feedback-Optimized Parallel Tempering Monte Carlo.” *Journal of Statistical Mechanics: Theory and Experiment*, **2006**(3), P03018. doi:10.1088/1742-5468/2006/03/p03018.
- Kitagawa G (1996). “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models.” *Journal of Computational and Graphical Statistics*, **5**(1), 1–25. doi:10.1080/10618600.1996.10474692.
- Kong A (1992). “A Note on Importance Sampling Using Standardized Weights.” *Technical Report 348*, The University of Chicago.
- Lakner C, Van der Mark P, Huelsenbeck JP, Larget B, Ronquist F (2008). “Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics.” *Systematic Biology*, **57**(1), 86–103. doi:10.1080/10635150801886156.
- Leiserson CE, Schardl TB, Sukha J (2012). “Deterministic Parallel Random-Number Generation for Dynamic-Multithreading Platforms.” *ACM SIGPLAN Notices*, **47**(8), 193–204. doi:10.1145/2370036.2145841.
- Lingenheil M, Denschlag R, Mathias G, Tavan P (2009). “Efficiency of Exchange Schemes in Replica Exchange.” *Chemical Physics Letters*, **478**(1–3), 80–84. doi:10.1016/j.cplett.2009.07.039.
- Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC. doi:10.1201/b13613.
- Lunn D, Spiegelhalter D, Thomas A, Best N (2009). “The BUGS Project: Evolution, Critique and Future Directions.” *Statistics in Medicine*, **28**(25), 3049–3067. doi:10.1002/sim.3680.
- Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000). “**WinBUGS** – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility.” *Statistics and Computing*, **10**(4), 325–337. doi:10.1023/a:1008929526011.
- Matsumoto M, Nishimura T (1998). “Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator.” *ACM Transactions on Modeling and Computer Simulation*, **8**(1), 3–30. doi:10.1145/272991.272995.

- Meng XL, Wong WH (1996). “Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration.” *Statistica Sinica*, **6**(4), 831–860.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953). “Equation of State Calculations by Fast Computing Machines.” *The Journal of Chemical Physics*, **21**(6), 1087–1092. doi:10.1063/1.1699114.
- Michail D, Kinable J, Naveh B, Sichi JV (2020). “**JGraphT** – A Java Library for Graph Data Structures and Algorithms.” *ACM Transactions on Mathematical Software*, **46**(2), 16. doi:10.1145/3381449.
- Milch B, Marthi B, Russell S, Sontag D, Ong DL, Kolobov A (2005). “BLOG: Probabilistic Models with Unknown Objects.” In *IJCAI: Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 1352–1359. Morgan Kaufmann Publishers.
- Mitchell TJ, Beauchamp JJ (1988). “Bayesian Variable Selection in Linear Regression.” *Journal of the American Statistical Association*, **83**(404), 1023–1032. doi:10.1080/01621459.1988.10478694.
- Murray LM, Schön TB (2018). “Automated Learning with a Probabilistic Programming Language: Birch.” *Annual Reviews in Control*, **46**, 29–43. doi:10.1016/j.arcontrol.2018.10.013.
- Neal RM (2001). “Annealed Importance Sampling.” *Statistics and Computing*, **11**(2), 125–139. doi:10.1023/a:1008923215028.
- Neal RM (2003). “Slice Sampling.” *The Annals of Statistics*, **31**(3), 705–741. doi:10.1214/aos/1056562461.
- Neal RM (2011). “MCMC Using Hamiltonian Dynamics.” *Handbook of Markov Chain Monte Carlo*, **2**(11). doi:10.1201/b10905.
- Ogata Y (1989). “A Monte Carlo Method for High Dimensional Integration.” *Numerische Mathematik*, **55**(2), 137–157. doi:10.1007/bf01406511.
- Okabe T, Kawata M, Okamoto Y, Mikami M (2001). “Replica-Exchange Monte Carlo Method for the Isobaric-Isothermal Ensemble.” *Chemical Physics Letters*, **335**(5–6), 435–439. doi:10.1016/s0009-2614(01)00055-0.
- Paige B, Wood F (2014). “A Compilation Target for Probabilistic Programming Languages.” arXiv:1403.0504 [cs.AI], URL <https://arxiv.org/abs/1403.0504>.
- Paige B, Wood F (2016). “Inference Networks for Sequential Monte Carlo in Graphical Models.” In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48 of *ICML’16*, pp. 3040–3049. New York.
- Plummer M (2003). “**JAGS**: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.” In K Hornik, F Leisch, A Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Technische Universität Wien, Vienna, Austria. URL <https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf>.

- Xtend (2019). “Xtend.” URL https://www.eclipse.org/xtend/documentation/203_xtend_expressions.html.
- Rathore N, Chopra M, de Pablo JJ (2005). “Optimal Allocation of Replicas in Parallel Tempering Simulations.” *The Journal of Chemical Physics*, **122**(2), 024111. doi:10.1063/1.1831273.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rodeh O (2008). “B-Trees, Shadowing, and Clones.” *ACM Transactions on Storage*, **3**(4), 2. doi:10.1145/1326542.1326544.
- Ronquist F, Kudlicka J, Senderov V, Borgström J, Lartillot N, Lundén D, Murray L, Schön TB, Broman D (2020). “Universal Probabilistic Programming Offers a Powerful Approach to Statistical Phylogenetics.” doi:10.1101/2020.06.16.154443. BioRxiv.
- Salvatier J, Wiecki TV, Fonnesbeck C (2016). “Probabilistic Programming in Python Using **PyMC3**.” *PeerJ Computer Science*, **2**(e55), 1–24. doi:10.7717/peerj-cs.55.
- Semmens BX, Ward EJ, Moore JW, Darimont CT (2009). “Quantifying Inter- and Intra-Population Niche Variability Using Hierarchical Bayesian Stable Isotope Mixing Models.” *PLoS ONE*, **4**(7), 1–9. doi:10.1371/journal.pone.0006187.
- Steele G, Lea D (2013). “Splittable Random Application Programming Interface.” URL <https://docs.oracle.com/javase/8/docs/api/java/util/SplittableRandom.html>.
- Steorts RC, Hall R, Fienberg SE (2016). “A Bayesian Approach to Graphical Record Linkage and Deduplication.” *Journal of the American Statistical Association*, **111**(516), 1660–1672. doi:10.1080/01621459.2015.1105807.
- Syed S, Bouchard-Côté A, Deligiannidis G, Doucet A (2019). “Non-Reversible Parallel Tempering: An Embarassingly Parallel MCMC Scheme.” arXiv:1905.02939 [stat.CO], URL <https://arxiv.org/abs/1905.02939>.
- Tancredi A, Liseo B (2011). “A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems.” *The Annals of Applied Statistics*, **5**(2B), 1553–1585. doi:10.1214/10-aos447.
- Teh YW, Jordan MI, Beal MJ, Blei DM (2006). “Hierarchical Dirichlet Processes.” *Journal of the American Statistical Association*, **101**(476), 1566–1581. doi:10.1198/016214506000000302.
- Van de Meent JW, Paige B, Yang H, Wood F (2018). “An Introduction to Probabilistic Programming.” arXiv:1809.10756 [stat.ML], URL <https://arxiv.org/abs/1809.10756>.
- van Rossum G, *et al.* (2011). *Python Programming Language*. URL <https://www.python.org/>.
- Wang L, Wang S, Bouchard-Côté A (2020). “An Annealed Sequential Monte Carlo Method for Bayesian Phylogenetics.” *Systematic Biology*, **69**(1), 155–183. doi:10.1093/sysbio/syz028.

- Wickham H (2014). “Tidy Data.” *Journal of Statistical Software*, **59**(10), 1–23. doi:10.18637/jss.v059.i10.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. 2nd edition. Springer-Verlag, New York. doi:10.1007/978-0-387-98141-3.
- Wickham H, François R, Henry L, Müller K (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7, URL <https://CRAN.R-project.org/package=dplyr>.
- Wood F, Van de Meent JW, Mansinghka V (2014). “A New Approach to Probabilistic Programming Inference.” In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pp. 1024–1032.
- Xie W, Lewis PO, Fan Y, Kuo L, Chen MH (2011). “Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection.” *Systematic Biology*, **60**(2), 150–160. doi:10.1093/sysbio/syq085.
- Zhao T, Cumberworth A, Wang Z, Gsponer J, de Freitas N, Bouchard-Côté A (2015). “Bayesian Analysis of Continuous Time Markov Chains with Application to Phylogenetic Modelling.” *Bayesian Analysis*, **11**(4), 1203–1237. doi:10.1214/15-ba982.
- Zhou Y, Gram-Hansen BJ, Kohn T, Rainforth T, Yang H, Wood F (2019). “LF-PPL: A Low-Level First Order Probabilistic Programming Language for Non-Differentiable Models.” In K Chaudhuri, M Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16–18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 148–157.
- Zhou Y, Johansen AM, Aston JAD (2016). “Toward Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach.” *Journal of Computational and Graphical Statistics*, **25**(3), 701–726. doi:10.1080/10618600.2015.1060885.

A. Advanced tutorials

A.1. Inference on non-standard data structures via third-party libraries

Consider an inference problem where the data structure or parameter of interest is a phylogenetic tree. A phylogenetic tree is a branching process encoding evolutionary relationships between organisms. The following example illustrates how to perform inference on a phylogenetic tree model given sequence alignment data, using a third-party library.

The Blang language itself does not contain tree-valued random variables. However, the language allows *creating* custom types of random variables. Moreover, these custom types can be packaged, published and imported.

First, we create a file called `dependencies.txt` at the root of the project directory. Each line in `dependencies.txt` encodes a versioned third-party library to be imported (along with its transitive set of dependencies). Here our model uses an existing Blang package called **conifer** providing phylogenetic-centric data types (Zhao, Cumberworth, Wang, Gsponer, de Freitas, and Bouchard-Côté 2015).⁷¹ Therefore, the contents for `dependencies.txt` would be given by the following.

```
com.github.UBC-Stat-ML:conifer:2.1.3
```

We encode the Blang model in `PhylogeneticTree.bl`, using the imported data type in Figure 34. In the first block of code, unobserved variables of the type `RealVar` and `IntVar` are declared using the functions `latentReal()` and `latentInt()`.⁷² This is no different from our previous examples. In the second block, `NonClockTreePrior` and `UnrootedTreeLikelihood` are themselves Blang models defined in the imported **conifer** package. `NonClockTreePrior` accepts a distribution as an argument, in the example an `XExpression` is used to pass in a Gamma distribution directly without the need to declare another variable in the model.

To run `PhylogeneticTree` we enter the following in the CLI:

```
$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example
$ blang --model jss.phylo.PhylogeneticTree \
> --model.observations.file data/primates.fasta \
> --model.observations.encoding DNA
```

```
Preprocess {
  ...
  Initialization {
    ...
  } [ ... ]
} [ ... ]
Inference {
  ...
  Round(9/9) {
```

⁷¹<https://github.com/UBC-Stat-ML/conifer/tree/master/src/main/java/conifer>

⁷²A summary of the most commonly used functions are listed in Figures 10 and 11.

```

package demo
import conifer.*
import static conifer.Utils.*

model PhylogeneticTree {

  random RealVar shape ? : latentReal()
  random RealVar rate ? : latentReal()
  random SequenceAlignment observations
  param EvolutionaryModel evoModel ? : kimura(observations.nSites)

  random UnrootedTree tree ? : unrootedTree(observations.observedTreeNodes)

  laws {
    shape ~ Exponential(1.0)
    rate ~ Exponential(1.0)
    tree | shape, rate ~ NonClockTreePrior(Gamma.distribution(shape, rate))
    observations | tree, evoModel ~ UnrootedTreeLikelihood(tree, evoModel)
  }
}

```

Figure 34: Phylogenetic tree model programmed in Blang, `PhylogeneticTree.bl`.

```

...
} [ ... ]
} [ ... ]
Postprocess {
  ...
} [ ... ]
executionMilliseconds : ...
outputFolder ../JSSBlangCode/.../results/all/2019-06-18-09-42-15-sP.exec

```

Here `--model.observations.file` specifies the data path, this is the standard Blang input method. `--model.observations.encoding` is a model-specific option to parse our data, provided by the third-party library. The usual outputs can be found in the `results` directory. On the whole, to use third-party libraries or packages (not necessarily restricted to Blang), users just need to specify the dependencies in `dependencies.txt`, and include `import` statements as needed. Running the model can be done via the usual CLI. Inputs follow the same syntax, unless otherwise instructed by the third-party library (i.e., custom parsers). Outputs are also placed in the usual directories.

A.2. Spike and slab classification

In this example, we focus on the implementation of a non-standard data type to handle a spike and slab model (Mitchell and Beauchamp 1988). The spike and slab model is a mixture of prior distributions commonly used for coefficients in a regression model. The non-standard data type `SpikedRealVar` is Blang's representation of the type of the coefficients in a spike and slab model. The file `SpikedRealVar.xtend` shown in Figure 35 contains the implementation of the data type `SpikedRealVar` using `Xtend`.

Because we want to use `RealVar` and `IntVar` types in our `SpikedRealVar` type (`Xtend`), we require the import statements of core Blang types, as the usual automatic imports are only

```

package jss.glm

import blang.core.RealVar
import blang.core.IntVar
import blang.types.StaticUtils

class SpikedRealVar implements RealVar {
    public val IntVar selected = StaticUtils::latentInt()
    public val RealVar continuousPart = StaticUtils::latentReal()

    override doubleValue() {
        if (selected.intValue < 0 || selected.intValue > 1)
            StaticUtils::invalidParameter()
        if (selected.intValue == 0) return 0.0
        else return continuousPart.doubleValue
    }
    override toString() { "" + doubleValue }
}

```

Figure 35: Spiked real variable programmed in Xtend, `SpikedRealVar.xtend`.

for Blang (`.bl`) files. We declare its member variables, `selected` and `continuous`, as `IntVar` and `RealVar`. These variables will encode the spike and slab component values for each explanatory variable. Because these members are random, their values are initialized using `latentInt()` and `latentReal()`. We override `RealVar()`'s getter method `doubleValue()` to return the regression coefficient if the explanatory variable is selected.

We can now use this custom data type to build a simple classification model, this time using Blang. The code in Figure 36 is contained in `SpikeSlabClassification.bl`.

In the above model, the random variable `parameters` is indexed by `instances` and `features`. This relationship is encoded using built-in types `Plated` and `Plate` variables; where a `Plated` variable is indexed by one or more `Plate` variable. Hence, the random variable `parameters` is of type `Plated<SpikedRealVar>`, and `instances` and `features` are of type `Plate<String>`. `Plate` and `Plated` variables are detailed in Section 10.4. Note this is not the only way to implement a spike and slab model. For instance, a user could define a distribution and sampler for the `SpikedRealVar` type itself. This example merely illustrates a minimal implementation that takes advantage of Blang's preexisting types, distributions, and samplers.

To perform inference on the model `SpikeSlabClassification`, we call the following using the CLI:

```

$ git clone https://github.com/UBC-Stat-ML/JSSBlangCode.git
$ cd JSSBlangCode/reproduction_material/example
$ blang --model jss.glm.SpikeSlabClassification \
> --model.data data/titanic/titanic-covariates.csv \
> --model.instances.name Name \
> --model.instances.maxSize 200 \
> --model.labels.dataSource data/titanic/titanic.csv \
> --model.labels.name Survived \
> --engine PT \
> --engine.nChains 20 \
> --engine.nScans 10000 \

```

```

package glms

model SpikeSlabClassification {

  param GlobalDataSource data
  random RealVar activeProbability ? : latentReal
  random RealVar sigma ? : latentReal
  random RealVar intercept ? : latentReal

  param Plate<String> instances, features
  param Plated<Double> covariates
  random Plated<IntVar> labels
  random Plated<SpikedRealVar> parameters

  laws {
    for (Index<String> instance : instances.indices) {
      labels.get(instance) | intercept,
      DotProduct dotProduct
      = DotProduct.of(features, parameters, covariates.slice(instance))
      ~ Bernoulli(logistic(intercept + dotProduct.compute))
    }

    for (Index<String> feature : features.indices) {
      parameters.get(feature).selected | activeProbability
      ~ Bernoulli(activeProbability)
      parameters.get(feature).continuousPart | sigma
      ~ StudentT(1.0, 0.0, sigma)
    }

    intercept | sigma ~ StudentT(1.0, 0.0, sigma)
    activeProbability ~ ContinuousUniform(0, 1)
    sigma ~ Exponential(1.0)
  }
}

```

Figure 36: Spike and slab model programmed in Blang, `SpikeAndSlabClassification.bl`.

```

> --postProcessor DefaultPostProcessor

Preprocess {
  ...
  Initialization {
    ...
  } [ ... ]
} [ ... ]
Inference {
  ...
  Round(9/9) {
    ...
  } [ ... ]
} [ ... ]
Postprocess {
  Post-processing activeProbability
  Post-processing allLogDensities
  Post-processing energy
  Post-processing intercept

```

index	features	mean	sd	min	median	max	HDI.lower	HDI.upper
1	Age	-0.02	0.02	-0.10	-0.02	0.01	-0.06	0
2	child	1.10	1.05	-1.10	0.96	5.23	-0.14	2.79
3	Fare	0.00	0.00	-0.02	0	0.02	0	0
4	female	3.16	0.44	1.43	3.15	4.72	2.40	3.85
5	Par..Aboard	0.01	0.13	-0.80	0	0.89	-0.21	0.26
6	Pclass	-0.51	0.29	-1.62	-0.52	0.16	-0.87	0
7	Sib..Aboard	-0.71	0.27	-1.90	-0.70	0.02	-1.15	-0.29

Table 4: Summary statistics for `SpikeSlabClassification` model’s parameters. Here 0.00 values are truncated representation of small values, and 0s denote exact 0 values as induced by the “spike” part of the model.

```

Post-processing logDensity
Post-processing parameters
Post-processing sigma
MC diagnostics
} [ ... ]
executionMilliseconds : ...
outputFolder ../JSSBlangCode/.../results/all/2019-06-18-10-05-29-ut.exec

```

The arguments `--model.data` and `--model.labels.dataSource` specify the data source, `--model.labels.name` and `--model.instances.name` specify the column names that `labels` and `instances` correspond to, and `--model.instances.maxSize` indicates the maximum size of the variable. The `--postProcessor` command creates additional summary statistics, posterior plots, trace plots, monitoring plots, and estimates of effective sample sizes in addition to the default outputs.

Figure 37 shows a subset of automatically post-processed trace plots. Summary statistics for `SpikeSlabClassification` model’s parameters are shown in Table 4.

B. Internal architecture

This section documents the high-level implementation decisions and trade-offs involved in the language construction. They may be skipped at first reading.

B.1. Language infrastructure

Blang is developed using Xtext, a mature framework for programming language design supported by the Eclipse Foundation and TypeFox. Thanks to the Xtext infrastructure, Blang incorporates a feature set comparable to many modern full-fledged multi-paradigm language: functional, generic and object programming, static typing. Blang also automatically inherits state-of-the-art language development tools including a graphical integrated development environment (IDE) which leverages static types to provide insight into large Blang projects. The IDE also has a full-feature debugger, and plug-ins have been tested to perform profiling and code coverage analysis.

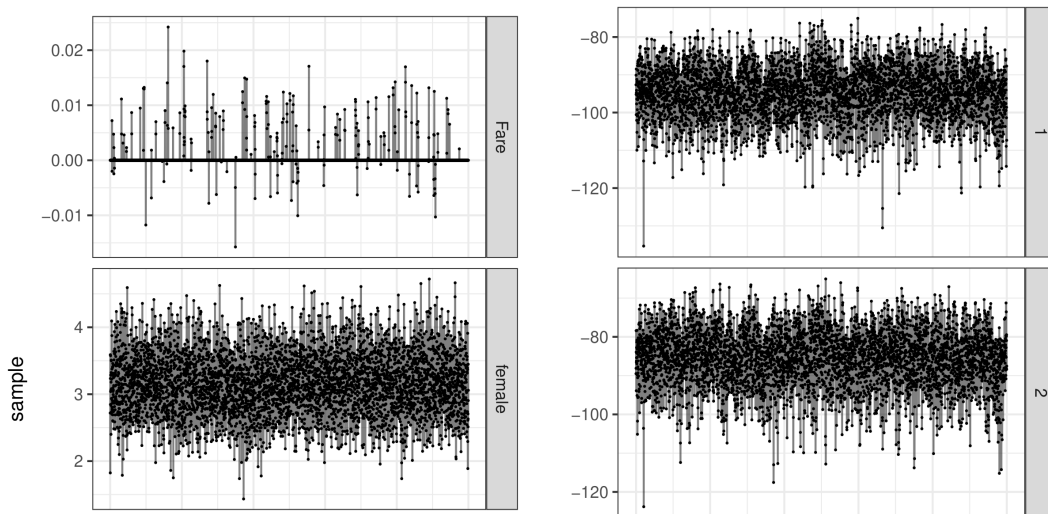


Figure 37: Facets of trace plots, produced by the default post-processor, for a subset of random variables in the spike and slab model. Note this is a truncated preview of the full set of trace plots, thus some axis labels and titles are not shown. For all plots, the x-axis indexes iteration. Left: the coefficients (y-axis) visit zero with positive probability as expected. Right: log densities (y-axis) for two of the 20 tempered chains used in PT.

B.2. Choice of compilation target

Under the hood, **Blang** is compiled into **Java**, which in turn is compiled into Java Virtual Machine (JVM) bytecode. This *transpilation* step does not have marked effect on amortized compilation time since we use compilers supporting incremental compilation.

This is the default model in **Xtext**, which, in addition to greatly simplifying **Xtext** development by using most of the provided default behavior, has for the user’s perspective two advantages related to performance and production deployment. First, code running on modern JVM is fast. For example, on the leading crowd-sourced language performance benchmark,⁷³ as of June 2019, the geometric mean performance of **Java** is lower than **C++**, but higher than **Julia** (Bezanson, Edelman, Karpinski, and Shah 2017), which itself outperforms the more common statistical computing choices such as **R** and **Python** by an order of magnitude or more. The performance gains of advanced compilers such as **Java** and **Julia** over **R** and **Python** are especially important when dealing with combinatorial spaces where vectorization is generally not possible. Other performance advantages include the JVM’s high-performance multi-threading capacity and garbage collection algorithms, which greatly facilitated the development of advanced Monte Carlo algorithms, for example for the parallel computation and memory management of particle genealogies. The second advantage is related to production deployment. **Java** is currently the most used language according to the TIOBE index as of June 2019, and this state may ease deployment of **Blang** software into existing production environments.

An often cited downside of using **Java** is its verbosity. In our context, one specific concern is that more boilerplate code is typically needed to access high-performance computing libraries

⁷³<https://benchmarksgame-team.pages.debian.net/benchmarksgame/which-programs-are-fastest.html>

such as linear algebra libraries or random number generators. Fortunately, **Blang** and **Xtend** avoid the key issues that make Java code verbose: checked exception, bad default behavior for constructor/accessors, and redundant type declaration. This brings **Blang** and **Xtend** code to a length similar to even non-statically typed language while preserving the advantages of the static type system. We use **Blang** and **Xtend** advanced language features combined with allowed operator overloading to wrap existing dense and sparse matrix libraries into **xlinear**, a new linear algebra library written in **Xtend**, which provides succinct linear algebra expressions to **Blang**. Similarly, we wrap existing random generation libraries into convenient **Xtend** extension methods.

B.3. Choice of sampler state representation

The state of the sampler is modified in place. A priori, this choice appears in conflict with a popular doctrine in software engineering which is to avoid mutability and instead use functional-style idioms on immutable data structures. While we agree these functional patterns are often tremendously helpful, in the context of our samples' state representation, we found mutable data structures more useful for three reasons. First, the way we precompute a factor graph for efficient inference, via scoping analysis, assumes that certain references in the object graph stay invariant. These invariant objects allow us to gain information on the scope and hence dependencies. With functional style programming, we would trade immutability of the values into more mutability of the references making this scoping analysis complex. Second, since the state objects are assembled and used in a completely automated way (via Java reflection), the user simply does not face the traps of mutable data structures in this specific context. Third, there are computational complexity advantages to using mutable data structures. For example, accessing or modifying entries in arrays has a cost of $O(1)$ (mutation) instead of $O(\log n)$ (copy-on-write) (Rodeh 2008).

C. Library dependencies

Blang's standard library uses its own language, and as such the majority of the dependencies were developed for **Blang**, and are handled automatically through **Maven** and **Gradle** (<https://gradle.org/>). Aside from libraries developed for **Blang**, **briefj**, **inits**, **bayonet**, **rejfree**, **binc**, **xlinear**, and **pxviz**,⁷⁴ the language depends on three additional, external libraries: **Cloning**,⁷⁵ **JGraphT** (Michail, Kinable, Naveh, and Sichi 2020),⁷⁶ and **Xbase** (Efftinge *et al.* 2013). For users who require automatic post-processors, R with packages **dplyr** and **ggplot2** is required. Table 5 summarizes each of the aforementioned packages, while the remainder of this section expands on a select few that have been referenced earlier in the paper.⁷⁷

C.1. bayonet

The **bayonet** library (<https://github.com/UBC-Stat-ML/bayonet>) contains utilities for performing probabilistic inference. **Blang** uses **bayonet.distribution.Random** as a replacement for

⁷⁴All of which are hosted on <https://github.com/UBC-Stat-ML/>.

⁷⁵<https://mvnrepository.com/artifact/uk.com.robust-it/cloning/1.9.6>

⁷⁶<https://mvnrepository.com/artifact/org.jgrapht/jgrapht-core/0.9.0>

⁷⁷An exhaustive list of dependencies used by the **blangSDK** package and their versions can be obtained by typing `./gradlew dependencies` from the root of the **blangSDK** directory.

Library	Description
briefj	Utilities for writing succinct Java code.
inits	A framework to organize inputs and outputs of scientific simulations.
bayonet	Various low-level utilities for probabilistic inference.
binc	An interface for calling binary programs from Java applications.
xlinear	Linear algebra package for Xtend and Java.
pxviz	A visualization library.
Cloning	Deep cloning library for Java.
JGraphT	Graph theory data structures and algorithms for optimizing samplers.
MathCommons	The Apache Commons Mathematics Library.
Xbase	Used as the base language for the DSL.

Table 5: A summary of library dependencies (automatically downloaded during installation, along with the transitive closure of these dependencies).

java.util.Random for random number generation. This alternative is compatible with both Java and **MathCommons** random types. **bayonet.math.SpecialFunctions** provides several statistical utility functions that are used heavily in Blang.

C.2. inits

inits (<https://github.com/UBC-Stat-ML/inits>) is a framework for performing scientific simulations, and can be viewed as a dependency injection framework tailored to complex and hierarchical command-line arguments. Blang’s CLI argument setup is automatically handled by **inits**.

C.3. xlinear

Blang’s linear algebra is based on **xlinear** (<https://github.com/UBC-Stat-ML/xlinear>), which itself relies on **ApacheCommons**, **ParallelColt**, and **JEigen**. The simple API of **xlinear** and the operator overloading functionality is what is leveraged in Blang to augment the **DenseMatrix** and **SparseMatrix** types into **DenseSimplex** and **DenseTransitionMatrix**.

D. Output format

Output organization

Every Blang execution creates a unique directory. The path is outputted to standard out at the end of the program’s execution/run. The latest run is also softlinked at **results/latest**. The directory has the following structure:

- **arguments-details.txt**: a detailed list of all arguments and options.
- **arguments.tsv**: arguments used in current run.
- **executionInfo**: information for reproducibility (JVM arguments, version of the code, standard out, etc.).

index	sample	value
0	0	0.453
1	0	0.386
2	0	0.886
0	1	0.520
1	1	0.345
2	1	0.940

Table 6: Example of the format used for posterior samples stored in tidy CSV files.

- `init`: information about the initialization process.
- `monitoring`: diagnostics for samplers.
- `samples`: samples from the target distribution. By default each random variable in the running model is output for each iteration (to disable this for some variables, e.g., those that are fully observed, use `--excludeFromOutput`).
- `logNormalizationEstimate.csv`: estimate of the natural logarithm of the probability of the data (also known as the log of the normalization constant of the prior times the likelihood, integrating over the latent variables).

Additional files and directories if `--postProcessor DefaultPostProcessor` is specified:

- `ess`: information for ESS and energy (negative log-likelihood) for each chain.
- `monitoringPlots`: sampler diagnostic plots.
- `posteriorPlots`: posterior densities and probability mass functions.
- `summaries`: summary statistics such as posterior means, HDIs, etc.
- `tracePlots`: trace plots for the random variables, log-density, and energy for each chain with burn-in samples discarded.
- `tracePlotsFull`: trace plots with all samples included.

Format of the samples

Posterior samples are stored in tidy CSV files. E.g., two samples for a `java.util.List` of three `RealVar`'s are shown in Table 6. By default, the method `toString` is used to create the last column (value). How can this be modified to encompass arbitrary data types? For example, how we output an object from permutation space (as in Section 9) as a tidy CSV is shown in Table 7.

This behavior can be customized to adhere to the tidy philosophy by implementing the interface `TidilySerializable` for a class of arbitrary data type.⁷⁸ The method `serialize`

⁷⁸<https://www.stat.ubc.ca/~bouchard/clang/javadoc-inits/clang/inits/experiments/tabwriters/TidilySerializable.html>

index	permutation_index	sample	value
0	0	0	2
0	1	0	0
0	2	0	1
1	0	0	1
⋮	⋮	⋮	⋮

Table 7: Example of the format used for posterior samples with arbitrary data types stored in tidy CSV files.

```

override void serialize(Context context) {
  for (int i : 0 ..< componentSize)
    context.recurse(connections.get(i), "permutation_index", i)
}

```

Figure 38: Permutation file `Permutation.xtend` used to instruct the sampler to parse and output the custom data type.

is invoked and passed an instance of `Context`.⁷⁹ Using `context.recurse(Object child, Object key, Object value)`, we can instruct the sampler to parse and output the custom data type as illustrated in Figure 38. The `child` argument is the value to write. The `key` is the name of the key, for example `permutation_index`. The `value` is the value of the key, for example `index i` of value in object.

Additional examples can be found in `TestTidySerializer.xtend`.⁸⁰

Output options

The following command-line arguments can be used to tune the output:

- `--excludeFromOutput`: space-separated list of random variables to exclude from output.
- `--experimentConfigs.managedExecutionFolder`: set to `false` in order to output in the current folder instead of in the unique folder created in `results/all`.
- `--experimentConfigs.recordExecutionInfo`: set to `false` to skip recording the reproducibility information in `executionInfo`.
- `--experimentConfigs.recordGitInfo`: set to `false` to skip git repository lookup for the code.
- `--experimentConfigs.saveStandardStreams`: set to `false` to skip recording the standard out and err.
- `--experimentConfigs.tabularWriter`: by default set to `CSV`. Can be set to `Spark` to organize tidy output into a hierarchy of directories each having a CSV (with less

⁷⁹See the `Context` design pattern.

⁸⁰<https://github.com/UBC-Stat-ML/inits/blob/master/src/test/java/blang/inits/TestTidySerializer.xtend>

columns, as many columns in this format can now be inferred from the names of the parent directories). In certain scenarios this could save disk space. Inter-operable with Spark.

E. List of probability distributions in Blang's library

E.1. Discrete distributions

Random variables in this section are integer-valued, hence `IntVars`. The distributions are listed in the following:

Bernoulli: Any random variable taking values in $\{0, 1\}$.

- `param RealVar probability`: Probability $p \in [0, 1]$ that the realization is one.

BetaBinomial: A sum of n IID Bernoulli variables, with a marginalized Beta prior on the success probability. Values in $\{0, 1, 2, \dots, n\}$.

- `param IntVar numberOfTrials`: The number n of Bernoulli variables being summed. $n > 0$.
- `param RealVar alpha`: Higher values bring the mean closer to one. $\alpha > 0$.
- `param RealVar beta`: Higher values bring the mean closer to zero. $\beta > 0$.

Binomial: A sum of n IID Bernoulli variables. Values in $\{0, 1, 2, \dots, n\}$.

- `param IntVar numberOfTrials`: The number n of Bernoulli variables being summed. $n > 0$.
- `param RealVar probabilityOfSuccess`: The parameter $p \in [0, 1]$ shared by all the Bernoulli variables (probability that they be equal to 1).

Categorical: Any random variable over a finite set $\{0, 1, 2, \dots, n - 1\}$.

- `param Simplex probabilities`: Vector of probabilities $(p_0, p_1, \dots, p_{n-1})$ for each of the n integers.

DiscreteUniform: Uniform random variable over the contiguous set of integers $\{m, m + 1, \dots, M - 1\}$.

- `param IntVar minInclusive`: The left point of the set (inclusive). $m \in (-\infty, M)$.
- `param IntVar maxExclusive`: The right point of the set (exclusive). $M \in (m, \infty)$.

Geometric: The number of unsuccessful Bernoulli trials until a success. Values in $\{0, 1, 2, \dots\}$.

- `param RealVar p`: The probability of success for each Bernoulli trial.

HyperGeometric: Hyper-geometric distribution with population N and population satisfying certain condition K and drawing n samples.

- `param IntVar numberOfDraws`: number of samples n .
- `param IntVar population`: number of population N .
- `param IntVar populationConditioned`: number of population satisfying condition K .

NegativeBinomial: Number of successes in a sequence of IID Bernoulli until (r) failures occur. Values in $\{0, 1, 2, \dots\}$.

- `param RealVar r`: Number of failures until experiment is stopped (generalized to the reals). $r > 0$.
- `param RealVar p`: Probability of success of each experiment. $p \in (0, 1)$.

Poisson: Poisson random variable. Values in $\{0, 1, 2, \dots\}$.

- `param RealVar mean`: Mean parameter λ . $\lambda > 0$.

YuleSimon: An exponential-geometric mixture.

- `param RealVar rho`: The rate of the mixing exponential distribution.

E.2. Continuous distributions

Random variables in this section are real-valued, hence `RealVars`. The distributions are listed in the following:

Beta: Beta random variable on the open interval $(0, 1)$.

- `param RealVar alpha`: Higher values bring the mean closer to one. $\alpha > 0$.
- `param RealVar beta`: Higher values bring the mean closer to zero. $\beta > 0$.

ChiSquared: Chi Squared random variable. Values in $(0, \infty)$.

- `param IntVar nu`: The degrees of freedom ν . $\nu > 0$.

ContinuousUniform: Uniform random variable over a close interval $[m, M]$.

- `param RealVar min`: The left end point m of the interval. $m \in (-\infty, M)$.
- `param RealVar max`: The right end point of the interval. $M \in (m, \infty)$.

Exponential: Exponential random variable. Values in $(0, \infty)$.

- `param RealVar rate`: The rate λ , inversely proportional to the mean. $\lambda > 0$.

F: The F-distribution. Also known as Fisher-Snedecor distribution. Values in $(0, \infty)$.

- `param RealVar d1, d2`: The degrees of freedom d_1 and d_2 . $d_1, d_2 > 0$.

Gamma: Gamma random variable. Values in $(0, \infty)$.

- `param RealVar shape`: The shape α is proportional to the mean and variance. $\alpha > 0$.
- `param RealVar rate`: The rate β is inverse proportional to the mean and quadratically inverse proportional to the variance. $\beta > 0$.

Gompertz: The Gompertz distribution. Values in $[0, \infty)$.

- `param RealVar shape`: The shape parameter ν . $\nu > 0$.
- `param RealVar scale`: The scale parameter b . $b > 0$.

Gumbel: The Gumbel distribution. Values in \mathbb{R} .

- `param RealVar location`: The location parameter μ . $\mu \in \mathbb{R}$.
- `param RealVar scale`: The scale parameter β . $\beta > 0$.

HalfStudentT: Half-Student T random variable. Values in $(0, \infty)$.

- `param RealVar nu`: A degree of freedom parameter ν . $\nu > 0$.
- `param RealVar sigma`: A scale parameter σ . $\sigma > 0$.

Laplace: The Laplace distribution over \mathbb{R} .

- `param RealVar location`: The mean parameter μ . $\mu \in \mathbb{R}$.
- `param RealVar scale`: The scale parameter b , equal to the square root of half of the variance. $b > 0$.

Logistic: A random variable with a logistic probability distribution function. Values in \mathbb{R} .

- `param RealVar location`: The center of the probability density function. Also the mean, mode and median. $\mu \in \mathbb{R}$.
- `param RealVar scale`: The scale parameter. $s > 0$.

LogLogistic: A random variable with a log-logistic probability distribution function. Values in $[0, \infty)$

- `param RealVar scale`: The scale parameter α and also the median. $\alpha > 0$.
- `param RealVar shape`: The shape parameter β . $\beta > 0$.

Normal: Normal random variables. Values in \mathbb{R} .

- `param RealVar mean`: Mean μ . $\mu \in \mathbb{R}$.
- `param RealVar variance`: Variance σ^2 . $\sigma^2 > 0$.

StudentT: Student T random variable. Values in \mathbb{R} .

- `param RealVar nu`: The degrees of freedom ν . $\nu > 0$.
- `param RealVar mu`: Location parameter μ . $\mu \in \mathbb{R}$.
- `param RealVar sigma`: Scale parameter σ . $\sigma > 0$.

Weibull: The Weibull distribution. Values in $(0, \infty)$.

- `param RealVar scale`: The scale parameter λ . $\lambda \in (0, \infty)$.
- `param RealVar shape`: The shape parameter k . $k \in (0, \infty)$.

E.3. Multivariate distributions

The distributions are listed in the following:

Dirichlet: The Dirichlet distribution over vectors of probabilities $(p_0, p_1, \dots, p_{n-1})$. $p_i \in (0, 1)$, $\sum_i p_i = 1$. Random variables with this distribution are of type **Simplex**.

- `param Matrix concentrations`: Vector $(\alpha_0, \alpha_1, \dots, \alpha_{n-1})$ such that increasing the i th component increases the mean of entry p_i .

MultivariateNormal: Arbitrary linear transformations of n IID standard normal random variables. Random variables with this distribution are of type **Matrix**.

- `param Matrix mean`: An $n \times 1$ vector μ . $\mu \in \mathbb{R}^n$
- `param CholeskyDecomposition precision`: Inverse covariance matrix Λ , a positive definite $n \times n$ matrix.

NormalField: A mean-zero normal, sparse-precision Markov random field. Random variables with this distribution are of type **Plated<RealVar>**.

- `param Precision precision`: Precision matrix structure.

SimplexUniform: n dimensional Dirichlet with all concentrations equal to one. Random variables with this distribution are of type **Simplex**.

- `param Integer dim`: The dimensionality n . $n > 0$

`SymmetricDirichlet`: n dimensional Dirichlet with all concentrations equal to $\frac{\alpha}{n}$. Random variables with this distribution are of type `Simplex`.

- `param Integer dim`: The dimensionality n . $n > 0$.
- `param RealVar concentration`: The shared concentration parameter α before normalization by the dimensionality. $\alpha > 0$.

E.4. Miscellaneous

`LogPotential`: A utility to handle undirected models (or random fields).

- `param RealVar logPotential`: The log of the current value of this potential.

F. Frequently used functions

Any Java function can be called in `Blang`. The functions in Figures 8 and 9 are automatically and statically imported for easy access. The functions provided in Table 8 are the most useful of those imported. In addition to these functions, `Blang` also imports two fields `E` and `PI` from `java.lang.Math`.

Function	Description
<code>abs(double value)</code>	absolute value
<code>acos(double a)</code>	arccosine
<code>asin(double a)</code>	arcsine
<code>atan(double a)</code>	arctangent
<code>cbrt(double a)</code>	cube root
<code>ceil(double a)</code>	ceiling
<code>cos(double a)</code>	cosine
<code>cosh(double a)</code>	hyperbolic cosine
<code>exp(double a)</code>	exponential base e
<code>floor(double a)</code>	floor
<code>log(double a)</code>	logarithm base e
<code>log10(double a)</code>	logarithm base 10
<code>max(double a, double b)</code>	maximum of a and b
<code>min(double a, double b)</code>	minimum of a and b
<code>pow(double a, double b)</code>	a to the power b
<code>signum(double a)</code>	signum function
<code>sin(double a)</code>	sine
<code>sinh(double a)</code>	hyperbolic sine
<code>sqrt(double a)</code>	square root
<code>tan(double a)</code>	tangent
<code>tanh(double a)</code>	hyperbolic tangent
Imported Fields	Description
<code>E</code>	Java's double value for e
<code>PI</code>	Java's double value for π

Table 8: Functions imported from `java.lang.Math`. Note that all trigonometric operations use angles expressed in radians and that the return type of all functions listed above are `double`.

Function	Description
<code>erf(double a)</code>	error function
<code>inverseErf(double a)</code>	inverse error function
<code>logistic(double a)</code>	standard logistic function
<code>logit(double a)</code>	standard logit function
<code>logBinomial(int n, int k)</code>	logarithm of $\binom{n}{k}$
<code>lnGamma(double alpha)</code>	logarithm of the gamma function of α
<code>logFactorial(int input)</code>	logarithm of the factorial of <code>input</code>
<code>multivariateLogGamma(int dim, double a)</code>	logarithm of the multivariate gamma function

Table 9: Functions imported from `bayonet.math.SpecialFunctions`. All return types are `double`.

Function	Description	Return Type
<code>latentInt()</code>	unobserved integer variable (initialized at zero)	<code>IntScalar</code>
<code>latentReal()</code>	unobserved real variable (represented as a double, initialized at zero)	<code>RealScalar</code>
<code>fixedInt(int value)</code>	fixed (constant or conditioned upon) integer scalar	<code>IntConstant</code>
<code>fixedReal(double value)</code>	fixed real scalar	<code>RealConstant</code>
<code>latentIntList(int size)</code>	size specifies the length of the list	<code>List<IntVar></code>
<code>latentRealList(int size)</code>	size specifies the length of the list	<code>List<RealVar></code>
<code>fixedIntList(int ... entries)</code>	list where the integer valued entries are fixed to the provided values	<code>List<IntVar></code>
<code>latentVector(int n)</code>	an n -by-1 latent dense vector (initialized at zero)	<code>DenseMatrix</code>
<code>fixedVector(double ... entries)</code>	an n -by-1 fixed dense vector	<code>DenseMatrix</code>

Table 10: Functions used to initialize random variables.

Function	Description	Return Type
<code>latentMatrix(int nRows, int nCols)</code>	an n -by- m latent dense matrix (initialized at zero)	<code>DenseMatrix</code>
<code>fixedMatrix(double [][] entries)</code>	a constant dense matrix	<code>DenseMatrix</code>
<code>latentSimplex(int n)</code>	latent n -by-1 matrix with entries summing to one (initialized at uniform)	<code>DenseSimplex</code>
<code>fixedSimplex(double ... probs)</code>	creates a constant simplex, also checks the provided list of number sums to one	<code>DenseSimplex</code>
<code>fixedSimplex(DenseMatrix probs)</code>	creates a constant simplex, also checks the provided vector sums to one	<code>DenseSimplex</code>
<code>latentTransitionMatrix(int nStates)</code>	latent n -by- n matrix with rows summing to one	<code>DenseTransitionMatrix</code>
<code>fixedTransitionMatrix(DenseMatrix probs)</code>	creates a constant transition matrix, also checks the provided rows all sum to one	<code>DenseTransitionMatrix</code>
<code>fixedTransitionMatrix(double [][] probs)</code>	creates a constant transition matrix, also checks the provided rows all sum to one.	<code>DenseTransitionMatrix</code>

Table 11: Frequently used functions in **Blang** to complement the set of functions from the **xlinear** library.

Function	Description	Return Type
<code>getRealVar(Matrix m, int row, int col)</code>	View a single entry of a <code>Matrix</code> as a <code>RealVar</code>	<code>RealVar</code>
<code>getRealVar(Matrix m, int index)</code>	View a single entry of a 1-by- n or n -by-1 <code>Matrix</code> as a <code>RealVar</code> .	<code>RealVar</code>
<code>asBool(int i)</code>	Returns false for 0 and true for 1	<code>boolean</code>
<code>asBool(IntVar i)</code>	Returns false for 0 and true for 1	<code>boolean</code>
<code>isBool(int i)</code>	Returns true if <code>i</code> is 0 or 1 and false otherwise.	<code>boolean</code>
<code>isBool(IntVar i)</code>	Returns true if <code>i</code> is 0 or 1 and false otherwise.	<code>boolean</code>
<code>asInt(boolean b)</code>	Returns 0 for false and 1 for true	<code>int</code>
<code>asList(Plated<T> plated, Plate<Integer> plate)</code>	Returns a the plated variable as a list.	<code>List<T></code>
<code>asCollection(Plated<T> plated, Plate<Integer> plate)</code>	Returns a the plated variable as a collection.	<code>Collection<T></code>
<code>asMap(Plated<T> plated, Plate<Integer> plate)</code>	Returns a the plated variable as a map.	<code>Map<K, T></code>
<code>generator(java.util.Random random)</code>	Upgrade a <code>java.util.Random</code> into <code>Random</code> to the type of <code>Random</code> <code>Blang</code> uses, <code>bayonet.distributions.Random</code> .	<code>Random</code>
<code>setTo(Matrix one, Matrix another)</code>	Copy the contents of a matrix into another one.	<code>void</code>
<code>sum(Iterable<? extends Number> numbers)</code>	Increment an entry of a map to double, setting to the value if the key is missing.	<code>void</code>
<code>increment(Map<T, Double> map, T key, double value)</code>	Check if two numbers are within $1e-6$ of each other.	<code>boolean</code>
<code>isClose(double n1, double n2)</code>		

Table 12: Extension methods (automatically) imported from `blang.types.ExtensionUtils`.

Function	Description	Return Type
<code>getName()</code>	Human-readable name for the plate, typically automatically extracted from a <code>DataSource</code> column name.	<code>ColumnName</code>
<code>indices(Query parentIndices)</code>	Get the indices available given the indices of the parent (enclosing) plates. The parents can be provided in any order.	<code>Collection<Index<K>></code>
<code>index(K key)</code>	Get the index given a key.	<code>Index<K></code>
<code>ofIntegers(ColumnName columnName, int size)</code>	a plate with indices 0, 1, 2, ... size-1	<code>Plate<Integer></code>
<code>ofStrings(ColumnName columnName, int size)</code>	a plate with indices category_0, category_1, ...	<code>Plate<String></code>
<code>ofStrings(String columnName, int size)</code>	a plate with indices category_0, category_1, ...	<code>Plate<String></code>
<code>get(Index<?> ... indices)</code>	get the random variable or parameter indexed by the provided indices. The indices can be given in any order.	<code>T</code>
<code>entries()</code>	list all variables obtained through <code>get(...)</code> so far. Each returned entry contains the variable as well as the associated indices (<code>Query</code>).	<code>Collection<Entry<Query, T>></code>
<code>slice(Index<?> ... indices)</code>	a view into a subset of the plated variable.	<code>Plated<T></code>
<code>latent(ColumnName name, Supplier<T> supplier)</code>	use the provided lambda expression to initialize several latent variables.	<code><T> Plated<T></code>

Table 13: Functions and methods related to `Plates` (above) and `Plated` (below) types.

Affiliation:

Alexandre Bouchard-Côté

Department of Statistics

Faculty of Science

University of British Columbia

3182 Earth Sciences Building, 2207 Main Mall Vancouver, BC Canada V6T 1Z4

E-mail: bouchard@stat.ubc.ca

URL: <https://www.stat.ubc.ca/~bouchard/>