Statistical Models

David M. Blei
Columbia University

October 25, 2016

We have discussed graphical models. Graphical models are a formalism for representing families of probability distributions. They are connected to efficient algorithms for computing marginals, conditionals, and other properties of the family. We now turn to the central question: **How can we use joint probability distributions to make meaningful statements about observed data?**

(A) In a statistical model, the data are represented as *observed random variables* $\mathbf{x}$. We set a *joint distribution* that also involves *hidden random variables* $p(\mathbf{z}, \mathbf{x})$. We calculate the *conditional distribution of the hidden variables given the observed variables* $p(\mathbf{z} \mid \mathbf{x})$. The conditional tells us about the data and helps form predictions.

(B) In a statistical model, the data are represented as *shaded nodes* in a *graphical model* that also involves *unshaded nodes*. We perform *probabilistic inference of the unshaded nodes*. This inference tells us about the data and helps form predictions.

Some examples:

- Observed: the audio signal;
  Hidden: uttered words.

- Observed: nucleotides;
  Hidden: whether they are associated with genes.

- Observed: words;
  Hidden: parts of speech.

- Observed: Facebook;
  Hidden: the overlapping communities within it.

- Observed: results of students in different classes;
  Hidden: a measure of the quality of each teacher.

- Others?

1 More Probability

We first go over a few more concepts from probability: continuous random variables, parameters, expectation and conditional expectation.

1.1 Continuous random variables

We have only used discrete random variables so far (e.g., dice). Random variables can be continuous. In a continuous random variable, we have a *density* $p(x)$, which integrates to one. If $x \in \mathbb{R}$ then

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{1}$$

For continuous random variables, probabilities are integrals over smaller intervals. For example,

$$P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x)dx.$$

Notice when we use P, p, X, and x.

Example: The Gaussian distribution. The Gaussian (or Normal) is a continuous distribution. Its density is

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2}$$

The density of a point x is proportional to the negative exponentiated half distance to μ scaled by σ^2. When x is closer to μ, its density is higher.

The Gaussian density is a bell-shaped bump,

[picture of a gaussian bump]

The variable μ is the *mean*; it controls the location of the bump. The variable σ^2 is the *variance*; it must be positive and it controls the spread around the bump.

1.2 Parameters

Parameters are values that *index* a family of distributions. As a simple example, a coin flip is a Bernoulli random variable. Its parameter is the probability of heads,

$$p(x \mid \pi) = \pi^{1[x=H]}(1-\pi)^{1[x=T]}. \tag{3}$$

(The notation $1[\cdot]$ is an *indicator function*: it is 1 when its argument is true and 0 otherwise.) Changing π leads to different instances of a Bernoulli distribution.

Gaussian. A Gaussian has two parameters, the mean and variance.

Beta. A beta random variable is over the $(0, 1)$ interval. Its density is

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{4}$$

Its parameters are α and β, both constrained to be positive. Note the function $\Gamma(\cdot)$ can be thought of as a real-valued generalization of the factorial function. For integers n, $\Gamma(n) = (n-1)!$.

The beta can have interesting shapes,

[pictures of beta random variables]

When $(\alpha, \beta) < 1$ it is like a bowl; when $(\alpha, \beta) > 1$ then it is a bump; when $\alpha = \beta = 1$ it is uniform. (You can see the uniform case from the equation.)

While we are looking at the beta, notice the following:

- The second term is

  $$\pi^{\alpha-1}(1-\pi)^{\beta-1}. \tag{5}$$

  It is a function of the random variable and parameters. The first term is

  $$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}. \tag{6}$$

  It is only a function of the parameters. It ensures that the density integrates to one; it is the *normalizing constant*. We saw the same quantity in undirected graphical models.

- In the Bernoulli the variable $\pi \in (0, 1)$ is the parameter. In the beta, it is the random variable. (We'll come back to this.)

Multinomial. Finally, a multinomial distribution (and categorical distribution) has a vector parameter on the k-*simplex*. (Technically, it is the $(k - 1)$-simplex but many authors use k-simplex.) The k-simplex is the space of non-negative vectors of length k that sum to one, i.e., $\theta_i > 0$ and $\sum_i \theta_i = 1$.

Dirichlet. The categorical is the vector generalization of the Bernoulli, indexed by a parameter in $[0, 1]$. The beta distribution is a distribution on $[0, 1]$. The vector generalization of the beta is the *Dirichlet distribution*. It is a distribution on the simplex.

A Dirichlet over the $k - 1$-simplex is parameterized by a k-vector $\alpha_j > 0$. The density is

$$\frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma\left(\alpha_j\right)} \prod_{j=1}^{k} \theta_j^{(\alpha_j - 1)} \tag{7}$$

Note that the beta is a distribution on the 1-simplex.

Discrete graphical models. What are the parameters to the kinds of discrete graphical models that we have been studying? They are the values of the conditional probability tables at each node. Each possible setting of tables leads to a specific distribution.

1.3 Expectation and conditional expectation

Consider a function of a random variable, $f(X)$. (Notice: $f(X)$ is also a random variable.) The expectation is a weighted average of f, where the weighting is determined by $p(x)$,

$$\mathbb{E}\left[f(X)\right] = \sum_x p(x) f(x). \tag{8}$$

In the continuous case, the expectation is an integral

$$\mathbb{E}\left[f(X)\right] = \int p(x) f(x) dx. \tag{9}$$

The conditional expectation is defined similarly

$$\mathbb{E}\left[f(X) \mid Y = y\right] = \sum_x p(x \mid y) f(x). \tag{10}$$

What is $\mathbb{E}\left[f(X) \mid Y = y\right]$? (A scalar.) What is $\mathbb{E}\left[f(X) \mid Y\right]$? (A random variable, i.e., a scalar for each value of y.)

The expectation is a function of the parameters. The expectation of a Gaussian random variable is

$$\mathbb{E}\left[X\right] = \mu.$$

The expectation of a beta variable is

$$\mathbb{E}\left[\pi\right] = \alpha/(\alpha + \beta).$$

Notice that it is in [0, 1]. (Not all expectations will have these simple forms.)

Expectations can provide useful summaries of distributions,

- Given someone's height, what is their expected weight?
- Given someone's demographic information, what is the expected number of dollars they will spend on your website.

Other expectations include higher moments of the distribution. For example, the variance is

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2. \tag{11}$$

This measures the spread of a distribution around its mean. In the Gaussian, $\mathrm{Var}\left[X\right] = \sigma^2$. In the beta it is

$$\mathrm{Var}\left[\pi\right] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{12}$$

Note how this matches the different shapes—bump shapes have lower variance, horn shapes have higher variance.

2 Statistical Models By Example

Now we can discuss statistical models. In a statistical model, the data are observed random variables. They are dependent on one or more hidden random variables.

For example, suppose n observations come from a Bernoulli with an unknown bias. This corresponds to the following graphical model,

[graphical model]

(Plates denote replication.) This is a very simple model, but it illustrates some of the ideas that frequently come into play:

- There is a hidden variable to represent what we don't know about the distribution of the data.
- Observations are "downstream" from the hidden variable. Information about the distribution "flows" to the observations.
- The data are conditionally independent given the hidden variable.

The graphical model naturally asks us to consider an inference about the unknown bias, the parameter (treated as a random variable) that governs the distribution of the data.

Let us work through this simple example.

Three ways to specify a model. There are three ways of specifying a statistical model. The first is by its graphical model.

But the graphical model does not tell us everything. The second way—and this way is fully specified—is by the joint distribution. The graphical model tells us the factorization of the joint. It is

$$p(\pi, x_{1:n}) = p(\pi) \prod_{i=1}^{n} p(x_i \mid \pi). \tag{13}$$

But we must specify the functional form of each of these terms. We know that $p(x_i \mid \pi)$ is a Bernoulli (Equation 3). We must also specify $p(\pi)$, a distribution over the parameter π. This variable is in (0, 1), so let's use a beta. For now we will hold $\alpha = \beta = 1$ to give a uniform beta variable (but we will get back to these so-called "hyperparameters" later.) We have fully specified the joint distribution.

A final way to think about a statistical model is through its *probabilistic generative process*. With the model we just defined, we have essentially described a program for generating data.

- Choose $\pi \sim \text{Beta}(\alpha, \beta)$.
- For each data point i:
  - Choose $x_i \mid \pi \sim \text{Bernoulli}(\pi)$.

We can easily write a program to execute this process. In essence, inference about the unknown bias π amounts to "reversing" the process to find the (distribution of the) the hidden variable that most likely generated the observations.

As an aside, writing models as programs is the core idea behind the field of *probabilistic programming*. In probabilistic programming we write a program that generates data from a model $p(\mathbf{z}, \mathbf{x})$. We then "compile" that program into a program that takes data as input $\mathbf{x}$ and returns samples (or a representation of) the posterior $p(\mathbf{z} \mid \mathbf{x})$. For example, this is the principle behind Stan (`http://mcstan.org`) and Edward (`http://edwardlib.org`).

Directed graphical models are particularly amenable to a generative process perspective. Because they are acyclic graphs, one can simply follow the flow of the nodes, iteratively generating each random variable conditional on its parents. As we mentioned above, information about the distribution of observations flows through the graph. Again, notice the theme of message passing. (Undirected graphical models do not lend themselves as easily to a stage-wise generative process.)

The posterior. When doing data analysis with a probability model, the central object of interest is the *the posterior*. The posterior is the conditional distribution of the hidden variables given the observations.[1] We have set up a model that describes how our data is generated from unobserved quantities. The posterior distribution describes the natural inference that one would want to make—what are the specific values of the hidden quantities that generated my data?

We can work through the posterior in this simple example. Encode $x = 0$ to mean

[1] Some definitions might differentiate between hidden variables that are traditionally seen as parameters and others that are traditionally seen as unobserved variables. But this is a fuzzy distinction. Let's not bother with it.

"tails" and $x = 1$ to mean "heads". The posterior is proportional to the joint,

$$\begin{aligned}
p(\pi \mid x_{1:n}) &\propto p(\pi) \prod_{i=1}^{n} p(x_i \mid \pi) && (14) \\
&= \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \right] \left[\prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} \right] && (15) \\
&\propto \left[\pi^{\alpha-1}(1-\pi)^{\beta-1} \right] \left[\pi^{\left(\sum_{i=1}^{n} x_i\right)}(1-\pi)^{\left(\sum_{i=1}^{n}(1-x_i)\right)} \right] && (16) \\
&= \pi^{\left(\alpha-1+\sum_{i=1}^{n} x_i\right)}(1-\pi)^{\left(\beta-1+\sum_{i=1}^{n}(1-x_i)\right)} && (17)
\end{aligned}$$

This is an unscaled beta distribution. It has the same functional form as the second term in Equation 4 with parameters $\hat{\alpha} = \alpha + \sum_{i=1}^{n} x_i$ and $\hat{\beta} = \beta + \sum_{i=1}^{n}(1 - x_i)$. Thus the posterior is $\text{Beta}(\hat{\alpha}, \hat{\beta})$.

Given a data set of n coin flips, this is the conditional distribution of the unknown bias. This kind of computation is the core of using statistical models to understand data.

Suppose our data truly come from a coin with a 70% chance of heads.

- Show the posterior with no observations
- Show the posterior with ten observations
- Show the posterior with twenty observations

2.1 Discussion

This is a good time to bring up a few interesting points.

Maximum likelihood estimation. First, let's momentarily veer from the probability model and consider a classical way of analyzing independent Bernoulli draws, the maximum likelihood estimate. The maximum likelihood estimate is the value of π that maximizes the probability of the data.

In this case, when we do maximum likelihood we assume the data are *independent and identically distributed* from a Bernoulli with unknown bias π. (Aside: In our graphical model are the x_i's IID?) We then estimate π to be the value for which the observed data was most probable. Typically, we work with the log probability, or log likelihood,

$$\begin{aligned}\hat{\pi}^{\mathrm{MLE}} &= \arg\max_{\pi} \log p(x_{1:n} \mid \pi) && (18)\\ &= \sum_{i=1}^{n} \log p(x_i \mid \pi) && (19)\\ &= \arg\max_{\pi} \sum_{i=1}^{n} \left(x_i \log \pi_i + (1 - x_i) \log(1 - \pi_i)\right). && (20)\end{aligned}$$

Optimizing with respect to the bias, subject to the constraint that it is in (0, 1), reveals that the MLE is the proportion of heads:

$$\hat{\pi}^{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{21}$$

When the parameter is part of a probability model, we are compelled to endow it with a so-called "prior distribution" $p(\pi)$, a distribution over the parameter that does not depend on data. (For example, above we chose a beta distribution.) This leads to estimates of its value that depend on the posterior. First, we discuss an important property of this example and its posterior.

Conjugacy. Notice that the posterior distribution is in the same family of distributions as the prior distribution, i.e., the distribution which we placed on the parameter. This is a property of pairs of distributions called *conjugacy*. The beta/Bernoulli are a conjugate pair. Conjugacy is an important concept throughout the rest of the course.

Conjugate pairs are useful. As you will see, we usually cannot compute the posterior exactly. However, *local* conjugacy—where individual pieces of a graphical model form conjugate pairs—will be important in approximate posterior inference.

Posterior expectation. We can calculate the expectation of the posterior beta, also known as the posterior expectation,

$$\mathbb{E}\left[\pi \mid x_{1:n}\right] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n}. \tag{22}$$

This is also called the *Bayes estimate* of the bias π. As n gets large, the Bayes estimate approaches the maximum likelihood estimate: the number of heads in the numerator and the number of trials in the denominator swallow the hyperparameters

α and β.

This is a property of Bayesian modeling. With fewer data, the prior plays more of a role in our estimate of the hidden variable; with more data it plays less of a role.

Also consider the posterior variance in Equation 12. The numerator grows as $O(n^2)$; the denominator grows as $O(n^3)$. Thus, as we see more data, the posterior becomes centered more closely to the MLE and becomes more "confident" about its estimate.

Warning: It is dangerous to overinterpret the posterior as providing an objective measure of "confidence" about the estimate. This posterior variance is a property of the model.

The posterior predictive distribution. Given our data, suppose we want to calculate the probability of the next coin flip being heads. What is it? Here we marginalize out the hidden variable,

$$\begin{aligned} p(x_{n+1} \mid x_{1:n}) &= \int p(x_{n+1}, \pi \mid x_{1:n}) d\pi && (23) \\ &= \int p(\pi \mid x_{1:n}) p(x_{n+1} \mid \pi) d\pi && (24) \\ &= \mathbb{E}\left[p(x_{n+1} \mid \pi) \mid x_{1:n}\right]. && (25) \end{aligned}$$

We see that this prediction depends on the posterior. It is a posterior expectation of a probability. Posterior inference is important both for learning about the parameter and for forming predictions about the future. (In this case, the posterior expected probability of x_{n+1} is the posterior expectation of the parameter itself.)

Penalized likelihood. Let's think more about the relationship between this kind of inference and maximum likelihood estimation.

We discussed one possible estimate, the posterior expectation. Another is the *maximum a posteriori* estimate, or MAP estimate. The MAP estimate is the value with maximum probability under the posterior. (Recall our discussion of inference on trees.)

Exploiting the log function again (a monotone function that preserves the maximum), the MAP estimate is found by solving

$$\begin{aligned} \hat{\pi}^{\text{MAP}} &= \arg\max_{\pi} \log p(\pi \mid x_{1:n}) && (26) \\ &= \arg\max_{\pi} \log p(\pi, x_{1:n}) && (27) \\ &= \arg\max_{\pi} \log p(\pi) + \sum_{i=1}^{n} \log p(x_i \mid \pi). && (28) \end{aligned}$$

The second line follows from the posterior being proportional to the joint. In the third line we see that the MAP estimate is found by solving a *penalized likelihood* problem. We will veer away from the MLE if it has low probability under the prior.

Note again what happens as we see more data. The prior term will be swallowed by the likelihood term, and the estimate will look more and more like the MLE.

3 Bayesian Statistics and Frequentist Statistics

We have discussed some basic ideas in probability, the general idea of statistical modeling, and a variety of ways to construe inference from data as a probabilistic computation. We demonstrated some important concepts, like conjugacy, that can simplify computation. More broadly, our goal is to use the tools of graphical models—tools that help us reason about probability models—to use models to compute about data. We now briefly discuss some more philosophical issues, namely what it means to be Bayesian and what it means to be frequentist.

Question for the class: Who is Bayesian? Who is a frequentist?

3.1 Bayesian Statistics

In Bayesian statistics, all inferences about unknown quantities are formed as probabilistic computations.

Suppose we have chosen the model structure, but not the parameters θ. For every setting of θ there is a different value of the probability of the data $x = x_{1:n}$, $p(x \mid \theta)$. As we have discussed, the goal of statistical inference is to “invert” this relationship, to learn about θ from x.

In Bayesian statistics, this happens through Bayes rule,

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}. \tag{29}$$

Discussion:

- We are treating θ as a *random variable*. The hallmark of Bayesian statistics is that all unknown quantities (or, rather, quantities that are not fixed) are random variables.

- As a consequence, we *have to* place a distribution on θ, the prior $p(\theta)$. To a true-blue Bayesian this should encode what we know about the parameters θ before seeing the data.[2] (I usually feel this is too a strong requirement and difficult in practical settings.)

- Our inference results in a *distribution* of the parameters given the data, $p(\theta \mid x)$, i.e., the posterior distribution. Computing the posterior is the central problem for Bayesian statistics.

3.2 Frequentist Statistics

Frequentists don't like to put priors on parameters and don't want to be restricted to probabilistic computations to estimate unknown quantities.

Rather, frequentists consider *estimators* for parameters θ and then try to understand the quality of those estimators using various theoretical frameworks and criteria. (Examples include *bias* and *variance*, which we discuss later when we talk about regression.)

While Bayesians condition on the data, frequentists treat the data as random and coming from an unknown distribution F. It is called "the population distribution." The estimator is a function of the data and thus it too is a random variable—its distribution depends on the population distribution. Frequentist criteria about the estimator are properties of its distribution.

For example, the most commonly used estimator is the maximum likelihood estimator. We treat $p(x \mid \theta)$ as a function of θ and then choose the value that maximizes this

[2] A friend of mine calls such statisticians "Crayesians".

function,

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} \log p(x \mid \theta) \tag{30}$$

(As usual, we take advantage of the monotonicity of log.) We can ask questions about $\hat{\theta}_{\mathrm{ML}}(x)$ such as its relationship to a "true" parameter θ^*, if x really came from the model, and its variance, which is a function of the number of data points n.

We can ask these questions because the estimator is a function of the data and thus a random variable. Its distribution is governed by the distribution of the data F.

Note that Bayesian computations can be thought of as estimators,

$$\begin{aligned} \hat{\theta}_{\mathrm{Bayes}}(x) &= \mathbb{E}\left[\theta \mid x\right] & (31) \\ \hat{\theta}_{\mathrm{MAP}}(x) &= \arg\max_{\theta} p(\theta \mid x) & (32) \end{aligned}$$

While these contain Bayesian models at the core, they can be analyzed with frequentist criteria. As we have discussed, in the latter case, the log posterior is called the penalized likelihood; the prior is called the regularizer.

And the lines between frequentist and Bayesian thinking can be blurred. Consider the problem of choosing the hyperparameters α. What if we were to choose them using maximum likelihood?

$$\begin{aligned} \hat{\alpha}_{\mathrm{ML}}(x) &= \arg\max_{\alpha} p(x \mid \alpha) & (33) \\ &= \arg\max_{\alpha} \int p(x \mid \theta) p(\theta \mid \alpha) d\theta. & (34) \end{aligned}$$

This style of computation is called *empirical Bayes*. It is a powerful idea.[3]

3.3 There is no debate

One perspective is that there is no real debate. More accurately, we must not confuse using Bayesian computation with being Bayesian. Being Bayesian means that all inferences about unknowns must arise through reasoning about a posterior, and there is no reason to assess these inferences in another way because we are Bayesian. Being frequentist means that we impose a theoretical framework on what it means to estimate an unknown, and then we evaluate our procedures (whether they come from

[3] And controversial: D. Lindley said: "There is nothing less Bayesian than empirical Bayes."

a posterior or another way) through that framework. As Freedman (1994) articulately points out, both schools of thought make assumptions and include models of the world.

Bayesian computation is computation about a posterior, regardless of *why* we think its a good idea. As examples, it may have good frequentist properties that we care about; our experience might indicate that it works well in practice; it might be the easiest way to analyze data that reflect assumptions we want to make (e.g., because we use a probabilistic programming system like Stan or Edward); or we might be Bayesians and will not consider another approach. In statistical machine learning we often lean on Bayesian computation because it is modular and convenient, rather than for philosophical reasons.

4 The Big Picture

We build a statistical model by developing a joint distribution of hidden and observed random variables. Our data are observed variables and the hidden quantities that govern them are hidden variables. In addition to the joint, a statistical model can be described by its graphical model or its generative process.

The key statistical quantity is the posterior, the conditional distribution of the hidden variables given the observations. The posterior is how we "learn" about our data via the model. The posterior predictive distribution gives us predictions about new data. Posterior estimates, like the posterior mean and the MAP estimate, tell us about the hidden variables that likely generated our data.

4.1 Example models

We will start getting used to thinking about models and also thinking about what their posterior gives us. When we design models, we design them to have useful posterior distributions. Here are some examples:

Mixture models. Mixture models cluster the data into groups. The hidden variables are the *cluster assignments* for each data point and the *cluster parameters* for each cluster. The cluster assignments form the grouping; the cluster parameters specify the distribution for the data in each group.

The posterior groups the data and describes each group. (Mixtures are the "simplest complicated model", and will form the core example when we discuss approximate posterior inference.)

Factor models. Factor models embed high-dimensional data in a low-dimensional space. The hidden variables are the per-data point component weights and the components. The component weights describe how each data point exhibits the components; the components themselves describe properties of the data that exhibit them. These models relate closely to ideas you have heard of like factor analysis, principal component analysis, and independent component analysis.

The posterior factorization describes a low-dimensional space (components) and each data point's position in it (weights).

Generalized linear models. *Regression models* or *generalized linear models* describe the relationship between input variables and response variables (or output variables). Each response variable (e.g., "income") arises from a distribution whose parameter is a weighted linear combination of the inputs (e.g., demographic information). The hidden variables are the weights of each type of input in the linear combination, called the coefficients. Related models, like hierarchical regressions and mixed-effect models, include additional hidden variables.

Regression models are called *discriminative* models because each data point is conditioned on inputs that are not modeled. The posterior of the coefficients describes a relationship between the inputs and the outputs. The posterior predictive distribution forms predictions from new inputs.

Here we note a distinction between "training" and "testing". When we fit the model we observe both inputs and response. When we use the model we only observe the inputs and must make inferences about the response.

This speaks to the larger distinction of unsupervised and supervised learning. In supervised learning, what we observe when fitting our model is different from what we observe when using it. In unsupervised learning, we never see the hidden variables. (As with all distinctions, many models are in a middle ground.)

Classification. Finally we can also examine a supervised "mixture" model, which is also known as *generative classification*. In this case the mixture assignments are known in training (e.g., spam and not spam). In testing, we infer the mixture assignments, i.e., classify new data points.

4.2 Example data (and corresponding models)

So far, we have described simple data sets and different types of models. Different types of *data* lead to further different types of models.

Dyadic data. Dyadic data are measured on pairs. In a social network, the data are edges between nodes; in web behavior, the data are which pages people clicked on.

For example, *matrix factorization* is a model of dyadic data.

Grouped data. Grouped data are data that are themselves individual data sets. Examples include test results from students grouped by class, documents that are groups of words, and images that are groups of image patches.

Hierarchical models capture grouped data. *Hierarchical regression models* are a supervised model of grouped data; *mixed-membership models* are an unsupervised model of grouped data.

Sequential and spatial data. *Sequential data* take place over time; *spatial data* are located in a space, such as latitude and longitude. Sequential data leads to dynamic models, like the hidden Markov model (HMM) or the Kalman filter (a linear dynamical model). Spatial data lead to spatial models, where correlation between data points relates to their distance.

In our discussion of the HMM, we assumed that the parameters were known. In some settings we can estimate these separately from observed data. For example, in speech recognition we can learn the audio model from observed utterances of words and the language model from written texts. But sometimes we cannot assume that they are known. In these unsupervised settings, inference involves learning both the state sequences (as before) as well as the transition matrices and observation probabilities.

4.3 Building models

I emphasize that these distinctions are blurry. Many models and data mix and match these properties, or express them to different degrees. There are many ways to construct latent spaces and to organize how data arises from them. Other ideas you might hear about include factorial models, Bayesian nonparametric models, mixed-effect models, multi-level models, and others.

The important point is that probabilistic modeling is not a cookbook of methods. Rather, these models can be used as modules in larger and more complicated models, customized to specific problems and applications. What is common in probabilistic modeling is that we treat our data as observed random variables in joint distribution that contains both observed and hidden quantities. We then use probabilistic computations about that model to help explore our data and form predictions.

Of course, we are conspicuously dancing around the important issue of "how do we choose a model?" This necessarily remains somewhat of a craft, the problem of translating your domain knowledge and knowledge of statistics into a graphical model that can help solve your problem.

Gelman et al. (1995) give a good summary of the process, another perspective on "Box's loop" (Blei, 2014).

1. Set up the full probability model, a joint distribution of all observed and hidden quantities. Try to write a model that is consistent with your knowledge about the underlying scientific problem and data collection process.

   (I add: write down a *sequence* of models from the least to more complicated. Consider an prioritization of the knowledge you want to inject in the model. Some is essential for solving a problem; others are instincts you have; others might be complications that, in retrospect, you do not need to address.

   And: think about the posterior, not necessarily the generative process. What unknown quantities do you hope to get out of the data?)

2. Conditioned on observed data, calculate and interpret the posterior distribution, the distribution of the unobserved variables of interest given the observed data.

3. *Evaluate* the fit of the model, and the implications of the posterior.

- Does the model fit the data?
- Are the conclusions reasonable?
- How sensitive are the results to the modeling assumptions?

If necessary, change the model.

There is no single organizing principle to organize statistical computations. But through probabilistic modeling, several distinctions emerge:

- **Bayesian versus Frequentist.** See above. Note that *both* perspectives involve probabilistic computations about the joint distribution.
- **Discriminative versus Generative.** Do we work conditional on some observations, or do we create a joint distribution about everything?
- **Per-data point prediction versus data set density estimation.** Do we care about predicting something about an unobserved quantity or are we more concerned about understanding the structure of the distribution that generated our observations?
- **Unsupervised versus supervised learning.** Is there an external signal that I care about? Do I observe it in one phase of my data, and is it hidden in the other phase?

Again, these are blurry boundaries. However, many statistical computations in all of these settings involve (a) treating observations as random variables in a structured probability model, and then (b) computing about that model.

References

Blei, D. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Freedman, D. (1994). Some issues on the foundation of statistics. *Foundations of Science*.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.