Hierarchical Regression

David M. Blei
Columbia University

December 3, 2014

Hierarchical models are a cornerstone of data analysis, especially with large grouped data. Another way to look at "big data" is that we have many related "little data" sets.

1 What is a hierarchical model?

There isn't a single authorative definition of a hierarchical model. Gelman et al. (1995) discuss two definitions:

1. "Estimating the population distribution of unonobserved parameters"
2. "Multiple parameters related by the structure of the problem"

Intuitively, knowing something about one "experiment" tells us something about another. For example:

- Multiple similar experiments
- Similar measurements from different locations
- Several tasks to perform on the same set of images

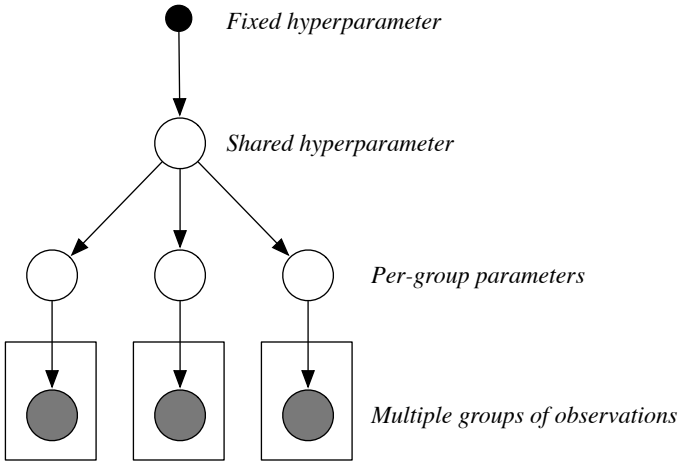
We've seen the last case when we talked about mixed-membership models. These are one type of hierarchical model.

When talking about hierarchical models, statiticians sometimes use the phrase "sharing statistical strength." The idea is that something we can infer well in one group of data can help us with something we cannot infer well in another.

For example, we may have a lot of data from California but much less data from Oregon. What we learn from California should help us learn in Oregon. The key idea is: *Inference about one unobserved quantity affects inference about another unobserved quantity.*

2 The classical hierarchical model

The classical hierarchical model looks like this:



We observe multiple groups of observations, each from its own parameter. The distribution of the parameters is shared by all the groups. It too has a distribution.

Inference about one group's parameter affects inference about another group's parameter. You can see this from the Bayes ball algorithm. Consider if the shared hyperparameter were fixed. This is *not* a hierarchical model.

As an intuitive example, consider height measurements of 8 year-old daughters from different families. Discussion:

- Clearly, which family you are in will affect your height.
- Assume each measurement is Gamma with an unknown per-family mean.
- Consider a family with 1000 kids. What do we know about the mean?
- How about when we observe 10 kids in a family? How about 1?
- What do we know about a new family that we haven't met?

Let's show mathematically how information is transmitted in the predictive distribution. Here is a more concrete (but still abstract) model:

- Assume m groups and each group has n_i observations.
- Assume fixed hyperparameters α.

Consider the following generative process—

- Draw $\theta \sim F_0(\alpha)$.
- For each group $i \in \{1, \ldots, m\}$:
  - Draw $\mu_i \mid \theta \sim F_1(\theta)$.
  - For each data point $j \in \{1, \ldots, n_j\}$:
    - Draw $x_{ij} \mid \mu_i \sim F_2(\mu_j)$.

For example, suppose all distributions are Gaussian and the data are real. Or, set F_0, F_1 and F_1, F_2 to be a chain of conjugate pairs of distributions.

Let's calculate the posterior distribution of the ith groups parameter given the data (including the data in the ith group). We are going to "collapse" the other means and hyperparameter. (But we are not worrying about computation. This is to make a mathematical point.) The posterior is

$$\begin{aligned} p(\mu_i \mid \mathcal{D}) &\propto p(\mu_i, \mathbf{x}_i, \mathbf{x}_{-i}) \\ &= \int p(\theta \mid \alpha) p(\mu_i \mid \theta) p(\mathbf{x}_i \mid \mu_i) \left(\prod_{j \neq i} \int p(\mu_j \mid \theta) p(\mathbf{x}_j \mid \mu_j) d\mu_j \right) d\theta \end{aligned}$$

Inside the integral, the first and third term together are equal to $p(\mathbf{x}_{-i}, \theta \mid \alpha)$. By the chain rule, this can be written

$$p(\mathbf{x}_{-i}, \theta \mid \alpha) = p(\mathbf{x}_{-i} \mid \alpha) p(\theta \mid \mathbf{x}_{-i}, \alpha). \tag{1}$$

Note that $p(\mathbf{x}_{-i} \mid \alpha)$ does not depend on θ or μ_i. So we can absorb it into the constant of proportionality. This reveals that the per-group posterior is

$$p(\mu_i \mid \mathcal{D}) \propto \int p(\theta \mid \mathbf{x}_{-i}, \alpha) p(\mu_i \mid \theta) p(\mathbf{x}_i \mid \mu_i) d\theta \tag{2}$$

In words, *the other data influence our inference about the parameter for group i through their effect on the distribution of the hyperparameter.* When the hyperparameter is fixed, the other groups do not influence our inference about the parameter for group i. (Of course, we saw this all from the Bayes ball algorithm.) Finally, notice that in contemplating this posterior, the group in question does not influence its idea of the hyperparameter.

Let's turn to *how* the other data influence θ. From our knowledge of graphical models, we know that it is through their parameters μ_j. We can also see it directly by unpacking the posterior $p(\theta \mid \mathbf{x}_{-i})$, the first term in Equation 2.

The posterior of the hyperparameter θ given all but the ith group is

$$p(\theta \mid \mathbf{x}_{-i}, \alpha) \propto p(\theta \mid \alpha) \prod_{j \neq i} \int p(\mu_j \mid \theta) p(\mathbf{x}_j \mid \mu_j) d\mu_j \tag{3}$$

Now we can make an interesting intuitive argument.

1. Consider each integral over μ_j in the second term as a weighted average of $p(\mu_j \mid \theta)$, weighted by $p(\mathbf{x}_j \mid \mu_j)$.

2. Suppose this is dominated by one of the values, call it $\hat{\mu}_j$. For example, there is one value of μ_j that explains a particular group $\mathbf{x}_j$.

3. Then we can approximate the posterior as

$$p(\theta \mid \mathbf{x}_{-i}, \alpha) \propto p(\theta \mid \alpha) \prod_{j \neq i} p(\hat{\mu}_j \mid \theta) \tag{4}$$

4. This looks a lot like a prior/likelihood set-up. If $p(\theta \mid \alpha)$ and $p(\hat{\mu}_j \mid \theta)$ form a conjugate pair then this is a conjugate posterior distribution of θ.

5. It suggests that *the hyperparameter is influenced by the groups of data through what each one says about its respective parameter.*

I emphasize that this is not "real" inference, but it illustrates how information is transmitted in the model.

- Other groups tell us something about their parameters.
- Those parameters tell us something about the hyperparameter.
- The hyperparameter tells us something about the group in question.

In Normal-Normal models, exact inference is possible. However, in general $p(\alpha \mid \mathcal{D})$ does not have a nice form.

Notice that the GM is always a tree. The problem—as we also saw with mixed-membership models—is the functional form of the relationships between nodes. In real approximate inference, you can imagine how an MCMC algorithm transmits information back and forth.

Consider other computations, conditioned on data.

- The predictive distribution for a known group $p(x_i^{\text{new}} \mid \mathcal{D})$ will depend on other groups (for the hyperparameter) and on other data within the same group (for the per-group parameter),

$$p(x_i^{\text{new}} \mid \mathbf{x}_i, \mathbf{x}_{-i}) = \int_\theta \left(\int_{\mu_i} p(x_i^{\text{new}} \mid \mu_i) p(\mu_i \mid \mathbf{x}_i, \theta) \right) p(\theta \mid \mathbf{x}_{-i}) \tag{5}$$

- The distribution of a parameter for a totally unobserved group μ^{new} will depend on the posterior of hyperparameter θ conditioned on all the data.

3 Hierarchical regression

One of the main application areas of hierarchical modeling is to regression and generalized linear models.

Hierarchical (or multilevel) modeling allows us to use regression on complex data sets. Some examples are:

- Grouped regression problems (nested structures)
- Overlapping grouped problems (non-nested structures)
- Problems with per-group coefficients
- Random effects models (more on that later)

Hierarchical/multi-level modeling is extremely flexible. Within the setting of linear modeling, it lets us use domain knowledge in many ways to capture how the response depends on the covariates. To see this, we will use two example problems.

The first is a recommendation problem.

- Echonest.net has massive music data, attributes about millions of songs.
- Imagine taking a data set of a user's likes and dislikes
- Can you predict what other songs he/she will like or dislike?
- This is the general problem of *collaborative filtering*.
- Note: We also have information about each artist.
- Note: There are millions of users.

A second (perhaps more typical) example is in the social sciences:

- We measured the literacy rate at a number of schools.
- Each school has (possibly) relevant attributes.
- What affects literacy? Which programs are effective?
- Note: Schools are divided into states. Each state has its own educational system with state-wide policies.
- Note: Schools can also be divided by their educational philosophy.

(Aside: The causal questions in the descriptive analysis are difficult to answer definitively. However we can, with caveats, still interpret regression parameters. That said, we won't talk about causation in this class. It's a can of worms.)

3.1 Notation, terminology, prediction & description

Quickly recall our notation and terminology about linear models

- *Response variable* y_n is what we try to predict.
- *Predictor vector* x_n are attributes of the ith data point.
- *Coefficients* β are parameters.

We discussed linear and logistic regression; we described how they are instances of generalized linear models. In both cases, the distribution of the response is governed by the linear combination of coefficients (top level) and predictors (specific for the ith data point). The coefficient component β_i tells us how the expected value of the response changes as a function of each attribute.

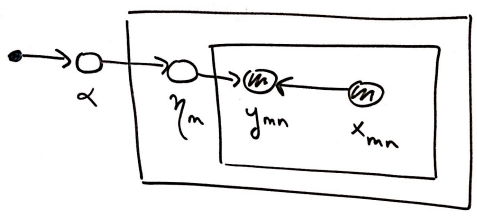
We can use regression for prediction or description.

- How does the literacy rate change as a function of the dollars spent on library books? Or, as a function of the teacher/student ratio?
- How does the probability of my liking a song change based on its tempo?
- Given a new song by an existing artist, will I like it?
- Given a new school, what is its expected literacy rate?

3.2 Hierarchical regression with nested data

The simplest hierarchical regression model simply applies the classical hierarchical model of grouped data to regression coefficients.

Here is the graphical model for nested regression:



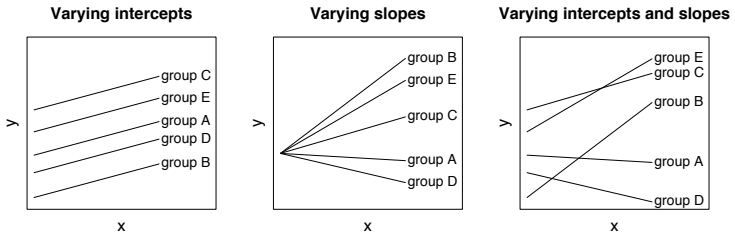
Here each group (i.e., school or user) has its own coefficients, drawn from a hyperparameter that describes general trends in the coefficients.

This allows for variance at the coefficient level. Note that the variance too is interpretable.

- A covariate can make a difference in different ways for different groups (high variance).
- A covariate can be universal in how it predicts the response (low variance).

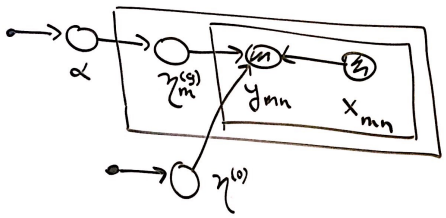
Further, because β is multivariate we have some flexibility about which components to model hierarchically and which components to model in a fixed way for all the groups. For example one distinction we make in regression is between the "slope" term and "intercept" term. The intercept term is (effectively) for a covariate that is always equal to one. (But intuitively we want to separate it from the rest of the coefficients, especially when regularizing.)

Gelman and Hill (2007) draw the following useful picture:



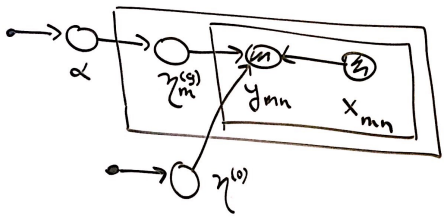
We hierarchically model the intercept, slope, or both coefficients.

The same choices apply for each cofficient or subset of coefficients. Often the problem will dictate where you want group-level coefficients or top-level coefficients. (Hierarchical regression means having to make more choices!)



We can also consider the noise term σ^2 to be group-specific. This let's us model errors that are correlated by group. Note that this affects inference about the hyperparameter θ. Higher noise at the group level means more posterior "uncertainty" about the group-specific coefficients. Those groups contribute less to the estimation of the hyperparameter.

Here is a more complicated nested regression:



Going back to the literacy example—

- Responses are grouped by state.
- Each state might have a different base literacy rate (intercept).
- The literacy rate increases or decreases in the same way as other schools, based on the attributes.
- Or, the relationship between some attributes and literacy rates is the same for all states; for others it varies across state.

Going back to the collaborative filtering example—

- The covariates are the artists of each song.
- Each artist has a different base probability that a user will like them. (Everyone loves Prince; not everyone loves John Zorn.) This is captured in the learned hyperparameter α.
- Some users are easier to predict than others. This is captured in per-group noise terms.

Recall the results from the previous section, and how they apply to regression. What I know about one school tells me about others; what I know about one user tells me about others. I can make a prediction about a new school or a new user, before observing data.

More extensions:

- Group-level attributes (like demographics of the user, or properties of the school) can be involved in both the response and the coefficients.

  When modeling coefficients, this gives us more power with respect to new users or new schools. Red-headed people might tend to like Phish. Montessori schools might tend to have higher literacy rates. These are statements about coefficients, conditional models of responses that we do not directly observe.

- Group-level coefficients can exhibit covariance. (This is not available in simple Bayesian models because there is only one "observation" from the prior.) For example, someone who likes Michael Jackson might be more likely to also like Janet Jackson. These correlations can be captured when we learn the hyperparameter α.
- We can nest models multiple layers deep.
- Any of the covariates (at any level) can be unobserved. These are called *random effects* and relate closely to matrix factorization, mixed-membership models, and mixture models.
- Really, hierarchical modeling bottoms out at using full power of plates, sequences, and directed graphical models on observations that are conditional on a set of covariates.

3.3 Regression with non-nested data

So far, we considered *grouped* or *nested* data. Hierarchical modeling can capture more complicated relationships between the hidden variables. Thus, it can capture more complicated relationships between the observations. (Alas, graphical models do not help us here.) Regression with non-nested data speaks to the "bottoming out" described above. Anything is possible (though inference might be tricky).

- Suppose each data point is associated with multiple groups encoded in g_n, a G vector with 1's for the group memberships.
- There are group-level parameters $\beta_{1:G}$, each a set of coefficients.
- Each observation y_n is drawn from a GLM whose coefficient is a sum of the coefficients of its groups

$$y_n \sim F\left(\left(\textstyle\sum_k \beta_k g_k\right)^\top x_n\right). \tag{6}$$

Note we don't *need* the per-group parameters to come from a random hyperparameter. They may exhibit sharing because data are members of multiple groups and even conditioning on the hyperparameter cannot untangle the data from each other.

That said, learning the hyperparameter will guarantee dependence across data. Nested data is an instance of this more general set-up.

Finally, there is an interesting space in between fully nested and fully non-nested models. This is where there are constraints to the multiple group memberships.

For example, consider again the recommendation problem. Each observation is a rating that a user i gives to a song j. While, in a sense, these observations belong to multiple groups (the user and the song), we also know that it cannot belong to two users or to two songs. Thus there is *overlapping nested structure*. In machine learning, these kinds of models are called factorization models. Two overlapping nested structures is a matrix factorization; more than that is a tensor factorization. We will discuss these at length later.

4 One word about inference

Stan.

References

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

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.