Hierarchical Modeling and Exchangeability (see also Bernardo and Smith, Ch.

Note: these notes are "work in progress", particularly the earlier material focussed on hierarchical modeling.

1 Hierarchical modelling

1.1 Examples:

In genomics there are many settings where we measure something on a large number of objects. A common example is measuring "gene expression" (which essentially means the activity level of a gene) on a large number of genes (e.g. approximately 20,000 genes in humans and chimps). Another example, is measuring the amount of transcription factor binding that occurs at a large number (thousands) of binding sites.

In other contexts it is common to measure the same thing on many examples (although perhaps not thousands). For example, we might measure the murder rate in many cities in the USA.

And in an example we looked at earlier, we considered having measured the DNA of several elephant tusks taken from a single "seizure", and using these measurements to infer whether each tusk came from a Savanna or a forest elephant.

The key idea that we want to get over here is that, in cases such as these, measurements on one object usually *actually give you information on the other objects.* Sometimes that information is very imprecise, sometimes it is very precise, but our first job is to convince ourselves that the general principle is, intuitively, true: measurements on some objects give information about other similar objects.

So consider the following examples:

- If you were told that the measured expression levels of 10 randomly-selected genes were 4.5, 6.2, 6.4, 7.5, 8.2, 6.3, 5.4, 10.1, 5.5 and 9.4, and were asked what you would guess for the expression level of the next gene, what would your answer be? Would it be different if you had instead seen 11.2,10.3,12.5,11.3,9.8,10.9,13.2,14.2,12.0, and 11.1?

- If I tell you that the murder rates in 5 randomly-selected US cities are 6.3,4.4,4.0,5.9, and 20.7 (per 100,000 people per year), what would you guess for the next one? Would it be different than if I'd told you 1.1,2.3,0.8,3.3 and 0.1?

- If you were testing a seizure of 100 tusks to determine their origins, and the first 95 all came from forest elephants. What would be your guess for the next one? Would it be different than if the first 95 had all come from savanna elephants?

The idea of sharing information across observations to improve inferences and estimates is sometimes referred to as "Borrowing Strength".

1.2 Hierarchical, or Multi-level, Modelling, to borrow strength

Hierarchical modeling provides a way, indeed the way, to capture the key idea above, and to "borrow strength" across observations.

Here's perhaps the simplest hierarchical model (related to the "normal means" problem).

Suppose you are measuring n related quantities, with error. Let $\beta_1, \ldots, \beta_n$ represent the true values, and $X_1, \ldots, X_n$ represent the measured values. Let $\sigma_1, \ldots, \sigma_n$ represent the standard deviations of each measurement, and we'll assume

these are known for now. We'll also assume that the errors are normal, so $X_i|\beta \sim N(\beta_i, \sigma_i^2)$. How should we estimate β_1 from the data?

The obvious estimate is $\hat{\beta}_1 := X_1$. But this does not borrow strength at all. It ignores the other observations!

Here's the idea: we can do better by assuming that the β_i come from some underlying (unknown) distribution. Then we can estimate this distribution from the data, which will allow us to borrow strength across observations.

For example, let's suppose we assume that $\beta_1, \ldots, \beta_n$ are independent and identically distributed $\sim N(\mu, \sigma^2)$. Of course we don't know μ and σ, but we can estimate them from the data. For example, the likelihood is given by:

$$L(\mu, \sigma) := p(X|\mu, \sigma) = \prod_i p(X_i|\mu, \sigma) \tag{1}$$

where $p(X_i|\mu, \sigma)$ is given by the normal density, $X_i \sim N(\mu, \sigma_i^2 + \sigma^2)$. We could estimate μ, σ by maximum likelihood.

Given μ and σ, we could estimate β_i using the posterior distribution $p(\beta_i|X, \hat{\mu}, \hat{\sigma})$. (Exercise: what is the posterior mean?)

This two-step procedure, using maximum likelihood to estimate the "hyper-parameters" μ and σ, is often called "Empirical Bayes".

Note that because we used maximum likelihood here for μ and σ, the posterior for β_i, $p(\beta_i|X, \hat{\mu}, \hat{\sigma})$, ignores the uncertainty in these "nuisance parameters", and will underrepresent the actual uncertainty in β. If the data are very informative for μ and σ then this is a small issue. If not then it is a bigger issue.

Of course, we could get around this by doing "full Bayes", specifying prior distributions for μ and σ, and integrating them out to obtain the posterior distributions for β.

Something to think about: when will this procedure work well? When might it work less well?

2 Exchangeability

Exchangeability is a mathematical concept that can motivate hierarchical modeling.

Definition: The random quantities $x_1, \ldots, x_n$ are said to be (finitely) **exchangeable** if

$$P(x_1 \in E_1, x_2 \in E_2, \ldots, x_n \in E_n) = P(x_{\pi(1)} \in E_1, \ldots, x_{\pi(n)} \in E_n)$$

for any permutation π of the set 1,2,...,n, and any (measureable) sets $E_1, \ldots, E_n$ of possible values. In terms of the corresponding pdf, this is equivalent to

$$p(x_1, \cdots, x_m) = p(x_{\pi(1)}, \cdots, x_{\pi(m)}).$$

An infinite sequence $x_1, x_2, \ldots$ is said to be exchangeable if every finite sequence is (finitely) exchangeable.

Intuitively, exchangeability is a statement that beliefs about the observeables $x_1, x_2, \ldots$ do not depend on the labelling or order in which we consider them.

Note that "independent and identically distributed" is a special case of exchangeable. (Simple Exercise.)

The assumption of exchangeability will be natural in many contexts. Canonical examples include tossing a coin, or a thumb-tack. More generally it will be natural when

1. The random quantities arise from a sample from a population (eg results of an opinion poll).
2. They arise as the result of an "experiment" which is repeated under similar conditions.

3. There is no natural labelling or indexing of the random quantities (for example because they do not arise in a time ordered way) and they are labelled for convenience.

Theorem (de Finetti): If $x_1, x_2, \ldots$ is an infinite exchangeable sequence of real valued random quantities with probability measure P, there exists a probability measure Q on $\mathcal{F}$, the set of all distributions on $\mathbf{R}$, such that the joint distribution of $x_1, \ldots, x_n$ has the form

$$P(x_1, \ldots, x_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(x_i) dQ(F)$$

Further,

$$Q = \lim_{n \to \infty} P(F_n) \tag{2}$$

where F_n is the empirical distribution function of $x_1, x_2, \ldots,$ (that is, F_n is the distribution which places mass n^{-1} on each observed value x_i, $i = 1, 2, \ldots$) and $P(F_n)$ denotes the law (distribution) of F_n.

Proof: See Bernardo and Smith, or most (advanced) textbooks on probability.

Remarks.

1. The theorem says that we can think of an exchangeable sequence $x_i, i = 1, 2, \ldots$ as arising from a "two-stage" randomization
   (a) First pick a distribution F according to the measure Q.
   (b) Conditional on F, $x_i, i = 1, 2, \ldots$ are i.i.d. F.
2. Consider the special case of an infinite sequence of {0,1}-valued random variables. A distribution on {0,1} is uniquely determined by the mass, p, on {1}. Thus the set of all distributions on {0,1} is effectively the unit interval [0,1]. The de Finetti measure Q is then just a probability distribution on [0,1]. In this context de Finetti's theorem says
   $$p(x_1, \ldots, x_n) = \int_0^1 p^{\Sigma x_i} (1-p)^{n - \Sigma x_i} dQ(p).$$
   In words, any infinite exchangeable sequence of {0,1}-valued random variables has a distribution of the form
   (a) Choose a value p for the success probability according to Q
   (b) Conditional on p, the x_i are i.i.d. Bernoulli(p).
   Further, $n^{-1} \sum_{i=1}^{n} x_i \to p$, so that p is the limiting long run proportion of successes.
3. Consider the case where the sequence $x_1, x_2, \ldots$ are either all 0s or all 1s, with equal probability. Clearly the values are not independent. But, they are exchangeable. Indeed, they are conditionally *iid* given p, where p is drawn from $\{0, 1\}$ with equal probability on each. [Note that this provides a nice simple illustration of Q being the limit of the empirical distribution. It also demonstrates that exchangeability does not imply i.i.d.].
4. An important consequence of de Finetti's theorem is that subjective beliefs which are consistent with (infinite) exchangeability *must* be of the form

(a) There are beliefs (*a priori*) on a parameter "F"

(b) Conditional on F the observations are i.i.d.

In general, the "parameter" F will be infinite dimensional. In particular contexts, additional structure in the beliefs may result in a restriction to finite dimensional parameter space. In this case, conventional terminology refers to the model as "parametric". In the infinite dimensional case it is often called "non-parametric".

5. In the presence of exchangeability, the "parameter" about which one has "prior" beliefs has a natural interpretation as a limit of observable quantities. (One concern sometimes expressed about Bayesian statistics is that it is "unnatural" to have to express "beliefs" about parameters. The point being made here is that in the presence of exchangeability, parameters are just functions of observables. For example in the exchangeable $\{0, 1\}$-valued setting, the "parameter" p is just the long-run proportion of successes, and it does not seem unreasonable to ask about uncertainty about this before seeing data.)

Exchangeability represents complete symmetry amongst the random quantities $x_1, x_2, \ldots$. There are many situations in which this might be unreasonable, but in which weaker symmetry conditions are plausible. These lead to versions of **partial exchangeability** for which analogies of de Finetti's theorem are available.

Note that even in situations where exchangeability does not formally apply, it can be a convenient and useful modeling approximation in practice (just as independence can be a convenient and useful modeling assumption). For example, if we are modeling data for the 50 states of the US, then the data are probably not formally exchangeable, but might be conveniently modeled as such.

Examples:

1. For $i = 1, 2, \ldots, L$, suppose that $x_{i1}, x_{i2}, \ldots$ represent measurements of components produced by the ith of L machines. For each i, it may be plausible that $x_{i1}, x_{i2}, \ldots$ are exchangeable, but this may not be plausible for x_{11}, x_{21}, x_{22} for example.

2. Drug Trial: In each of I contexts (e.g. type of subject with regard to age, sex, ... in a drug trial) there may be J different treatments (e.g. drugs) applied to each of K subjects. Writing x_{ijk} for the response of the kth subject given treatment j in context i, ($i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$) it may be reasonable to assume that x_{ij1} and x_{ij2} are exchangeable for each i and j, but exchangeability across i and j may be implausible.

3. Consider a specific case of Example 1 above in which the x_{ij} are {0,1} valued. Under the assumption that $x_{i1}, x_{i2}, \ldots$ are exchangeable for each i, an analogue of de Finetti's theorem shows that the joint distribution of the x_{ij} must have the form

$$\begin{aligned} p(x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{L1}, \ldots, x_{Ln_L}) &= \int_{[0,1]} \theta_1^{\Sigma x_{1j}} (1-\theta_1)^{n_1 - \Sigma x_{1j}} \\ &\quad \ldots \theta_L^{\Sigma x_{Lj}} (1-\theta_L)^{n_L - \Sigma x_{Lj}} dQ(\theta_1, \ldots, \theta_L), \end{aligned}$$

where $\theta_i = \lim_{n\to\infty}(x_{i1} + \cdots + x_{in})/n$, for $i = 1, 2, \ldots, L$.

That is, the joint distribution is as if there were unknown success probabilities θ_i associated with the ith component, and conditional on $(\theta_1, \ldots, \theta_L)$, the x_{ij} are independent Bernoulli r.v.'s with

$$P(x_{ij} = 1 \mid \theta_1, \ldots, \theta_L) = \theta_i, \quad j = 1, 2, \ldots.$$

If the machines are "similar", then it may be natural to regard the L long run success frequencies $\theta_1, \ldots, \theta_L$ as exchangeable. If further, the L machines may be thought of as a choice from a potentially infinite collection, then we may regard $\theta_1, \ldots, \theta_L$ as part of an infinite exchangeable sequence. In this case, de Finetti's theorem gives

$$Q(\theta_1, \ldots, \theta_L) = \int \prod_{i=1}^{L} G(\theta_i) d\pi(G).$$

That is, it is *as if* there were a random distribution G on [0,1], and conditional on G, the θ's are i.i.d. G. This gives a heirarchichal model for the data $x_{i1}, x_{i2}, \ldots$, $i = 1, 2, \ldots$.

Finite Exchangeability

Throughout this subsection we have considered infinite exchangeable sequences. Suppose instead that we have a finite collection $x_1, \ldots, x_n$ which we regard as exchangeable. We can use de Finetti's theorem *if* we can regard $x_1, \ldots, x_n$ as part of an infinite collection $x_1, x_2, \ldots$. Formally, we need the existence of an exchangeable distribution for the infinite sequence whose restriction to the first n elements coincides with our distribution for $x_1, \ldots, x_n$.

1. Not all finite exchangeable sequences can be embedded in an infinite sequence in this way. For a simple example, x_1, x_2 with
$$P(x_1 = 1, x_2 = 0) = P(x_1 = 0, x_2 = 1) = \frac{1}{2},$$
cannot be embedded in an exchangeable sequence x_1, x_2, x_3. (Exercise)

2. If $x_1, \ldots, x_n$ can be embedded in an exchangeable $x_1, \ldots, x_N$ for large N, then (in an appropriate sense) the distribution of $x_1, \ldots, x_n$ will be close to that given by the de Finetti representation. (Diaconis and Freedman, Ann. Prob. 8 (1980) 745-764.)

3 Example: Multiple Linear Regression

Consider the multiple linear regression model,

$$Y = \beta X + \epsilon \tag{3}$$

where Y is an (observed) n-vector of responses, X is an (observed) $n \times p$ matrix of predictors, β is a p vector of effects (to be estimated), and $\epsilon \sim N(0, 1/\tau)$ is an (unobserved) n-vector of errors. Assume that both Y and X have been "centered" to have 0 mean in each column (this is a common trick that is used so that there is no need to worry about an intercept).

If p is large compared with n, then maximum likelihood estimation of β will not work (too many parameters leads to an underdetermined system).

Solution: "borrow strength" on the β values using a hierarchical model.

3.1 Non-sparse version

Simplest version of borrowing strength is to assume

$$\beta_j | \tau, \sigma_b \sim N(0, \sigma_b^2/\tau). \tag{4}$$

Note 1: we could assume a non-zero mean, but it is usually not appropriate. For example, you would usually want procedures to be invariant to recoding a column of X as $-X$, which requires a symmetric prior for β about 0.

Note 2: the scaling of the prior by τ is standard: it simplifies computation, and makes results transform appropriately with changes in units of Y (e.g. changing Y from feet to inches would scale estimates of β by a factor of 12).

In an empirical Bayes approach, we would estimate hyper-parameters σ_b^2 and τ by maximum likelihood. (The likelihood is easy, integrating out β: exercise). And the posterior distribution on $\beta | \sigma_b, \tau, Y, X$ is available in closed form.

This procedure always leads to estimated β values that are closer to 0 than the least squares estimates (e.g. when these exist, $p < n$). This effect is referred to as "shrinkage", and the prior (4) is referred to as a "shrinkage prior". [Frequentists like shrinkage too: it can reduce variance, at the cost of some bias, resulting in a smaller mean squared error.]

3.2 Sparse version

The assumption that the β values are normally-distributed is a little restrictive. Various deviations from normality could be considered, but a common one is to allow that some values of β might be exactly zero (or so close as to be indistinguishable), while the remainder are normally distributed. This leads to the following prior:

$$p(\beta | \pi, \sigma_b, \tau) = (1 - \pi)\delta_0 + \pi N(0, \sigma_b^2/\tau) \tag{5}$$

where δ_0 denotes a point mass on zero.

With this prior, integrating out to obtain a likelihood for π, τ and σ_b is no longer easy. In practice people have either fixed π, σ_b to "default" values (estimating τ), or performed a full Bayesian analysis, placing priors on these unknowns and estimating them by computational methods to be discussed (MCMC).

Note that the resulting model can encourage β to be sparse (if π is near 0), and learns how sparse by borrowing strength across all β values to learn π. Alternative approaches to encouraging sparsity include penalized regression such as the Lasso and Elastic net.

Further reading: Y Guan and M Stephens. Bayesian Variable Selection Regression for Genome-wide Association Studies, and other Large-Scale Problems. Annals of Applied Statistics. 5(3): 1780-1815, September 2011.