Hierarchical Linear Models and Cell Data

Geoffrey J. Gordon
ggordon@cs.cmu.edu

March, 2000

Abstract

Hierarchical linear models are a generalization of Bayesian linear regression. They differ from Bayesian regression in that they determine the width of the prior distribution for the regression coefficients automatically from the data. They are particularly appropriate when we want to answer questions about the typical weights in a regression instead of just the typical examples. This paper provides a tutorial on hierarchical linear models, then demonstrates their application to some biochemical data.

1 Introduction

Consider an experiment in which we fertilize half of each of three fields, then measure the height of several plants grown in each half-field. To analyze our measurements from this experiment, we might run a linear regression in which we try to predict the heights of the plants from which field they were in and whether they were fertilized.

In this regression, the three regressors for the effects of the different fields are a priori interchangeable: we wouldn't be able to tell if someone swapped the measurements from two of the fields. In addition, we expect that any new field we measured would be interchangeable with the old ones. In other words, we believe that there is some distribution of fields, and that each field we measure is a new sample from that distribution.

Hierarchical linear models are a generalization of linear regression which makes this view explicit. They assume that the true weights are independent samples from some underlying distribution, and they try to estimate this distribution explicitly.

There are several reasons we might want to fit a hierarchical model instead of a standard regression. First, the number of regressors might depend on how many samples we collect. Hierarchical models provide a convenient way to say that each new regressor will be similar to some of the old ones. Second, we might not have a good idea how large the true weights are, but we might expect that they are all about the same size. Conventional regression can't express this belief, but a hierarchical model can. Finally, we might want to answer questions

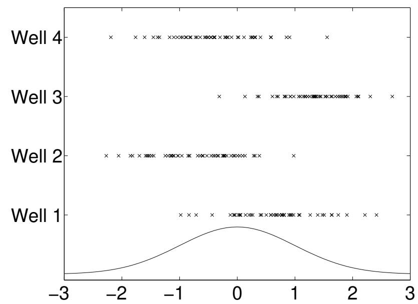

Figure 1: Two possible distributions for the measured outputs of cells. In one panel cell readings are nearly independent, while in the other cells in the same well are correlated. The marginal distribution of cell readings is the same in both panels, so if we fail to take account of the correlations within wells we may make incorrect inferences.

about the distribution of groups of regression coefficients. Conventional regression doesn't estimate this distribution, and so can only answer such questions indirectly.

The first description of Bayesian hierarchical models was [1], but that paper uses a biased estimator for variances. A better treatment of variances and regularization is in [2]. A good but dense overview is [3]. Hierarchical models are related to complexity control methods for neural nets, particularly the Automatic Relevance Determination model.

2 A biochemical problem

When a pharmaceutical company discovers a new compound that might affect synthesis of an important protein, it conducts an experiment like the following one to assess the compound's value.[1] A researcher grows cultures of cells on a glass plate. The cultures are separated by raised barriers into regions called wells. The researcher prepares solutions of the target compound at various concentrations and applies one solution to each culture. After a specified delay, she exposes the cells to a dye that binds the protein of interest and examines several randomly selected cells from each well under a microscope. She records how well the dye bound to each cell, then repeats the experiment with new plates and different concentrations of the solution.

In the simplest case there are just two doses of interest, zero and high or control and treatment, and we want to answer the question of whether there is a difference between treated and untreated cells. One way to answer this question is to separate the treatment and control groups, compute the mean and variance of the dye levels for each group, and compare the means under a normal model.

[1] We have changed the description of our data slightly because of disclosure requirements.

	wells					plates		
	1	2	3	4		1	2	
cell 1	1					1		
cell 2	1					1		
cell 3		1				1		
cell 4		1				1		
cell 5			1		...		1	...
cell 6			1				1	
cell 7				1			1	
cell 8				1			1	

Figure 2: Part of the design matrix for an ANOVA.

There is a problem with this procedure, though: our data are not independent. Cells in the same well or wells on the same plate will be correlated because there are unmeasured influences that are constant within a well or plate. For example, an error in titrating the solution for a well affects all cells within that well, while a change in the amount of time the plate is incubated before dyeing affects all wells on that plate.

Figure 1 shows an example of this problem. The two panels show two possible distributions for the dye levels of 200 cells in 4 wells exposed to the same concentration of the drug. In the top panel there is significant correlation between cells in the same well, while in the bottom panel each cell's reading is approximately independent. The best estimate of the mean dye level μ is $\hat{\mu}$, the average of all 200 observations. If we ignore correlations within wells, we can estimate $\mathrm{Var}(\hat{\mu})$ with the sample variance divided by the number of samples. Since the observed variance is about 1 in either panel, the estimated standard error is about 0.071.

In the bottom panel this estimate is almost right since the cells are nearly independent. (The actual standard error is 0.086.) But in the top panel we are ignoring a significant source of variance: while we can estimate the well means accurately with 50 samples apiece, there are only 4 samples we can use to infer the distribution of well means. Since well-level variation contributes significantly to the overall variance, our uncertainty in the well distribution dominates our overall uncertainty. So, the actual standard error of $\hat{\mu}$ in the top panel is 0.357.

3 Analysis of variance

The classical technique of analysis of variance, or ANOVA, solves part of the problem of accounting for correlations within groups of cells. An ANOVA is just a linear regression with a real output and binary inputs. In our case the output is the dye level and the inputs describe which well each cell is in, whether it is a treatment or a control, and so forth.

In more detail, we can construct a matrix X called the design matrix. Each row of X represents a cell. Each column of X represents a binary feature of the cells. For example, one column might be 1 if cell i is in well j and 0 otherwise, while another might be 1 if cell i is on plate k. Figure 2 shows part of our design matrix.

Write y for the output vector, so that y_i is the dyed-ness of cell i. We can solve $y \approx Xw$ in the least squares sense for the weight vector w. Each element of w tells us how much one feature affects the output on average, controlling for all other features. For example, w_j might tell us how much more brightly dyed the cells in well j are than we would have expected from knowing what plate they're on, whether they're controls, and so forth. From the outcome of the regression, hypothesis tests can answer questions like "are treated cells more brightly dyed on average than controls?"

4 The problem with ANOVA

Often in an analysis of variance we want to ask a question about a group of weights. For example, we might want to know the fraction of total variability due to variation between wells.

This sort of question only makes sense if the weights form a natural group. The property that lets us group weights is called exchangeability. Two weights w_1 and w_2 are exchangeable if for any numbers x and y we have no prior preference for $w_1 = x$, $w_2 = y$ over $w_1 = y$, $w_2 = x$. See [5, ch. 5] for a readable explanation of exchangeable parameters in hierarchical models. Weights we decide to group together are called random effects; weights we treat separately are called fixed effects.

The problem with analysis of variance is that it is designed for fixed effects only: it treats each weight identically and does not take into account any possible grouping. It is possible to hack ANOVA to answer questions about groups of weights, but as we will see there are logical problems with asking about group properties of fixed effects, so the answers that ANOVA gives us do not always make sense.

The obvious way to estimate the variance of a group of weights with ANOVA is to estimate each weight, then add up the squares of the estimates and divide by the number of weights. Unfortunately, this approach doesn't account for the uncertainty in our estimate of w. This uncertainty causes an upward bias in the sum of squared weights. Worse, this bias may not go to zero as our sample size increases: if the number of weights grows with the number of samples, as it does in our cell data, there will be a minimum amount of uncertainty that is always present in our estimate of w.

We can try to correct for our uncertainty in w by calculating our expected bias and subtracting it off. This is the estimator which is usually recommended for random effects in a classical ANOVA. Unfortunately, it can predict a negative variance, even in large samples. To fix this problem, we need to turn to a Bayesian analysis.

5 Bayesian hierarchical linear models

Classical regression fits the model $y \sim N(Xw, \sigma^2 I)$, which states that the outputs are normally distributed with equal variances σ^2 and with the mean of y_i equal to $X_i \cdot w$. Here X_i is the ith row of the design matrix X. This model does not let us express any prior information we may have about w. Bayesian regression allows us to provide a prior distribution for w; here we will assume that this prior is normal with diagonal covariance, $w_i \sim N(0, \alpha_i^2)$ for some constant vector α.

Bayesian hierarchical linear models take one step further: instead of requiring that we know the variances α_i^2 ahead of time, they estimate some or all of these variances from the data. They are called "hierarchical" because they conceptually separate inference into two levels: inferring the weights w from X and y, and inferring the variances α_i^2 from w. Of course, these levels cannot be completely separated, since α influences the inference of w as well as vice versa.

Since there are as many α_is as w_is, we need constraints to make α well determined. If we can divide the weights into exchangeable groups, the natural constraints are that $\alpha_i = \alpha_j$ if w_i and w_j are in the same group. With these constraints each group of weights follows a normal distribution with zero mean and unknown variance. In other words, all weights in a group are likely to be about the same size, but we don't know what size.

We can now state more simply the difference between fixed and random effects: fixed effects have fully-specified prior distributions, while priors for random effects contain parameters to be estimated. With this definition, it is clear why we can't ask what the data tell us about the variance of a new fixed effect: the variance of a new fixed effect is a property of the prior, and the prior is fully specified before seeing any data.

It may seem like cheating not to specify our prior for w until after we see the data, but it is not. A hierarchical model specifies a distribution for y just as a traditional linear model does. The difference is that in the traditional model w and σ are the parameters we need to estimate from the data, while in the hierarchical model the parameters are σ and α. w is not a parameter in the sense of a root cause which is not influenced by other variables; it is a hidden variable, a random quantity which we do not observe but which influences the observed data and is influenced by the parameters.

In a hierarchical model we can generate a sample from the joint prior distribution of w and y given X, σ, and α by first picking each w_i as $N(0, \alpha_i^2)$, then picking y as $N(Xw, \sigma^2 I)$. If A is the diagonal matrix with $A_{ii} = \alpha_i^2$, then the marginal distribution of the observed data y after integrating out the hidden variable w is $N(0, \sigma^2 I + XAX^{\mathrm{T}})$.[2]

[2]The conclusion that $E(y) = 0$ looks like a problematic limitation. There are two answers to this objection. First, $E(y)$ is only zero if we have no knowledge of w. Once we have seen some data, we no longer believe $E(w) = 0$, so we no longer believe $E(y) = 0$. Second, even before seeing any data, we can fix some of the α_is at large or infinite values to reduce or eliminate the shrinkage on y. This trick is often useful in combination with a redefinition of X (*e.g.*, to include a new parameter for the mean of a group of weights).

A fully Bayesian treatment would specify prior distributions for α and σ. To keep our notation simple we will assume that these priors have negligible weight.[3]

6 Algorithms

As mentioned above, a hierarchical model has two kinds of unknowns: the weights w and the variances σ^2 and α_i^2. If we knew α ahead of time we could find the posterior distribution for w and σ by Bayesian regression. With α unknown, though, estimation becomes more complicated. The problem is that the natural parameter of a zero-mean Gaussian distribution is the inverse covariance matrix, also called the precision. The estimation equations have a simple form when expressed in terms of the natural parameter, but in our case the precision is $(XAX^{\mathrm{T}} + \sigma^2 I)^{-1}$, which is an essentially nonlinear function of α_i^2 and σ^2. So, we need an iterative algorithm to solve for σ and α.

We can divide algorithms for hierarchical models into two classes: deterministic methods such as EM, and Monte Carlo methods such as Gibbs sampling. Monte Carlo methods are generally slower, but they permit us to compute the full joint posterior distribution for w, σ, and α. Deterministic methods can be faster, but they generally use point estimates for σ and α.

In large samples, the variances α_i^2 and σ^2 will usually be well determined. So, it will not be crucial to find their exact posterior distribution, and a point estimate such as the one computed by EM will be sufficient. The weight vector w, on the other hand, may not be well determined even in large samples: for example, if there are only a few cells per well, the well means will always be poorly determined no matter how many cells we measure. The joint posterior distribution of w and the variances has a skewed peak, so the values of w, σ, and α that jointly maximize the likelihood will be biased estimates, even in the limit of infinite sample size. This is why it is essential that EM provides a full posterior for w and not just a point estimate.

In smaller samples the posterior distribution for α and σ will have significant spread. In this case we have two options: we can continue to use iterative methods along with the formulas given in [3] for error bars on the variances, or we can switch to Monte Carlo methods. For small enough samples it will probably be necessary to switch, but for our experiments the EM algorithm was sufficient.

[3] If the assumption of negligible priors is a problem, it is simple to implement a conjugate prior for these parameters: in each M step of the EM algorithm as described below in Section 7, instead of setting a variance to the corrected sum of squares divided by the number of contributions, we can set it to the corrected sum of squares plus a constant, divided by the number of contributions plus another constant. The two constants together define the prior, and they can be chosen separately for each variance parameter.

7 Expectation-maximization

This section describes the EM algorithm. For a different iterative algorithm see [3]; for the Gibbs sampling algorithm see [5].

The EM algorithm is a recipe for estimating the parameters of a distribution when some data are hidden. It alternates between two steps: for fixed parameters, compute the expected log likelihood of the observed data by integrating out the hidden data. Then find the new parameter values which maximize the expected log likelihood, and repeat. The two steps are called E and M for expectation and maximization. EM is useful because it is often easier to compute the expected log likelihood than the expected likelihood. (The latter is what we would need in order to maximize the likelihood directly.) A theorem guarantees that each cycle of EM improves the expected likelihood unless the parameters are already at a stationary point [6].

In our case the input matrix X and output vector y are observed, the weights w are hidden, and σ and α are the parameters we need to estimate. Recall that A is the matrix with α_i^2 on the diagonal. The joint log likelihood of y and w is

$$\begin{aligned}-2\ln P(y,w|X,\sigma,A) &= \text{constant} + \\ &N\ln\sigma^2 + \sigma^{-2}\|y - Xw\|^2 + \ln\det A + w^{\mathrm{T}}A^{-1}w\end{aligned}$$

For the E step we need the posterior distribution of w for fixed A and σ. Since the log likelihood is quadratic in w, the posterior distribution of w is Gaussian. Collecting the terms containing w and completing the square shows that the precision is $B = A^{-1} + \sigma^{-2}X^{\mathrm{T}}X$ and the mean is $\beta = \sigma^{-2}B^{-1}X^{\mathrm{T}}y$. So, to find the expected log likelihood we just need to take the expectation of a quadratic function of a Gaussian random variable. The result is, up to additive and multiplicative constants,

$$\begin{aligned}&N\ln\sigma^2 + \sigma^{-2}(\|y - X\beta\|^2 + E(\|X(w-\beta)\|^2)) \\ &\qquad + \ln\det A + E(ww^{\mathrm{T}})\circ A^{-1}\end{aligned}$$

Here we have written $X \circ Y$ to mean the "dot product" of the two equal-size matrices X and Y, that is, the sum of the products of corresponding elements.[4] To obtain this expression we rewrote $w^{\mathrm{T}}A^{-1}w$ as $ww^{\mathrm{T}} \circ A^{-1}$ and split $E\|y - Xw\|^2$ into mean^2 + variance or $\|y - X\beta\|^2 + E(\|X(w-\beta)\|^2)$.

We can now split $E(ww^{\mathrm{T}})$ into mean^2 + variance or $\beta\beta^{\mathrm{T}} + B^{-1}$. Also, the expression $E(\|X(w-\beta)\|^2)$ is the sum of the variances of the elements of Xw. Since the variance of w is B^{-1}, the variance of Xw is $XB^{-1}X^{\mathrm{T}}$. So, the sum of the variances of the elements of Xw is the trace of $XB^{-1}X^{\mathrm{T}}$, or $XB^{-1}X^{\mathrm{T}} \circ I = B^{-1} \circ X^{\mathrm{T}}X$. So, we can write the expected log likelihood as

$$N\ln\sigma^2 + \sigma^{-2}(\|y - X\beta\|^2 + B^{-1}\circ X^{\mathrm{T}}X)$$

[4] Some authors write trace(XY) instead of $X \circ Y$. We use the latter notation to emphasize that $X \circ Y$ is linear in X or Y. A useful identity is

$$X^{\mathrm{T}}AX \circ B = A \circ XBX^{\mathrm{T}}$$

for any X, A, B. If w is a vector, we have the special case $w^{\mathrm{T}}Aw = ww^{\mathrm{T}} \circ A$.

$$+ \ln\det A + B^{-1} \circ A^{-1} + \beta\beta^{\mathrm{T}} \circ A^{-1}$$

At this point we are done with the E step. To perform the M step we need to maximize the above expression over σ and α. A possible point of confusion is that σ and α appear in the definition of B; but, since B is a parameter of the distribution of w, the EM algorithm specifies that we should hold B fixed (using the values of σ and α from the previous M step) while performing the current M step.

Since we are holding B fixed, the log likelihood splits into two pieces, one containing only σ and one containing only α. So to perform the M step we can solve for σ and α separately. Differentiating with respect to σ^{-2} and setting to zero gives

$$N\sigma^2 = \|y - X\beta\|^2 + B^{-1} \circ X^{\mathrm{T}}X$$

In other words, to estimate σ^2 we take the sum of squared residual errors, add a little bit to compensate for the fact that we optimized w, and divide by the number of observations.

To solve for α we need some extra notation. Write b for the diagonal of the matrix $B^{-1} + \beta\beta^{\mathrm{T}}$, so that $b_i = E(w_i^2)$. Write ρ_k^2 for the variance of the kth group of weights, so that $\alpha_i = \rho_k$ if weight i is in group k. Finally, write N_k for the number of weights in group k. Now, $\ln\det A = \sum_i \ln \alpha_i^2$; so, collecting terms containing ρ_k gives

$$N_k \ln \rho_k^2 + \sum_{i \in k} b_i \rho_k^{-2}$$

where the sum runs over all indices i belonging to weights from group k. Taking the derivative with respect to ρ_k^{-2} and setting to zero yields

$$N_k \rho_k^2 = \sum_{i \in k} b_i$$

In other words, to estimate the variance of group k, we sum up the expected squared weights in group k and divide by the size of the group.

The above equations for σ and α form the M step of the EM algorithm.

8 EM details

In many cases the precision B is a sparse matrix. (This will often happen if the design matrix X contains several large groups of non-overlapping indicator features, as is true in the cell experiment described below.) Unfortunately, B^{-1} will not normally be sparse, so we would like to avoid representing B^{-1} explicitly.

To avoid working directly with B^{-1}, we need to reorder computations in two places. The first place is in the expression $B^{-1} \circ X^{\mathrm{T}}X$; the second is when we are computing the diagonal of B^{-1} so that we can find the vector b of expected squared weights. In both cases we can save work by replacing B with a factorization. In our experiments we used the Cholesky factorization $U^{\mathrm{T}}U$ where U is upper triangular. (Both U and U^{-1} were sparse. We did not need to preorder B before factoring it, although for other problems with different patterns of sparsity in the design matrix X, a preordering or even a different factorization might work better.)

To compute $B^{-1} \circ X^{\mathrm{T}}X$ we can rewrite it as $XB^{-1}X^{\mathrm{T}} \circ I$. Since $B^{-1} = U^{-1}U^{-\mathrm{T}}$, this is the sum of the diagonal elements of $XU^{-1}U^{-\mathrm{T}}X^{\mathrm{T}}$ or the sum of the squared lengths of the rows of XU^{-1}. To compute the i, ith element of B^{-1}, we took the squared length of the ith row of U^{-1}.

We used the biased estimator described in Section 4 as a starting point for EM.

9 An example

An interesting and informative special case of the EM algorithm for Bayesian hierarchical linear models is when we are estimating the mean and variance of a normally distributed sample y_i of size N. In this case there is only one regression parameter, the mean μ, so we will fix its prior variance at infinity. There remains one variance parameter to estimate, namely σ^2, the residual variance. Solving for the fixed point of the EM iteration from Section 7 shows that the posterior precision of μ is $B = N/\sigma^2$, and

$$\begin{aligned} N\sigma^2 &= \sum_i (y_i - \mu)^2 + B^{-1}N \\ (N-1)\sigma^2 &= \sum_i (y_i - \mu)^2 \end{aligned}$$

which is the classical unbiased estimator of the residual variance.

10 Experiments

We analyzed 12,657 cells distributed over 96 wells on 2 plates. Each well was exposed to one of 12 concentrations of the drug; there were between 77 and 163 cells measured per well, with an average of about 132.

We want to know how many cells we will need to measure in future experiments to say confidently whether the drug is present. (The number of cells needed will depend on the minimum concentration we will be required to detect.) To answer this question we need to estimate several quantities. First, we need the dose-response curve, that is, the expected change in dye level at each concentration of the drug. This curve tells us how big a response to expect and therefore how precisely we need to measure it in order to separate it from zero. Second, we need the well-to-well variability. This variance tells us how many different wells we need to sample from to cancel out well-level noise. Finally, we need the residual variance to tell us how many cells we need to measure to achieve a given level of accuracy.

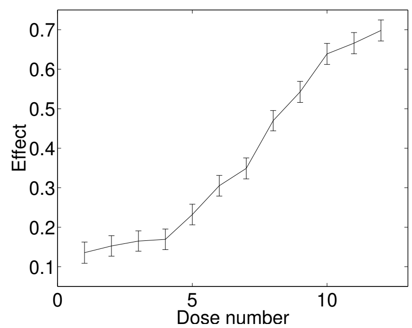

Figure 3: Dose-response curve. Horizontal axis shows our 12 drug concentrations from smallest to largest. Vertical axis shows expected change in dye level.

To estimate these quantities we set up a hierarchical linear model. The model's design matrix has $110 = 96 + 12 + 2$ columns, containing indicator variables for each well, concentration, and plate. It is rank-deficient (14 null eigenvalues) since the 12 concentrations and 2 plates span a subspace of the well effects. We fixed the prior variances for the plate effects at infinity; we then assigned the 96 well effects to one exchangeable group and the 12 concentration effects to another. This setup causes the plate effects to capture the overall mean, forcing the well and concentration effects to sum to zero.

The concentration effects are not truly exchangeable, since we expect higher concentrations of the drug to cause a greater response. We modeled them as a group anyway since we wanted to convey the prior information that they were likely to be of similar sizes. In order for the concentration effects to be exchangeable, the drug doses would need to be chosen independently rather than spaced out, and we would need to hide the smallest-to-largest ordering of the concentrations. We believe that the effect of these two violations of exchangeability is minor.

10.1 EM results

Fitting our hierarchical model by EM gives a well-to-well standard deviation of $\alpha = 0.0273$ and a residual standard deviation of $\sigma = 0.302$. The dose-response curve is shown in Figure 3.[5]

The EM algorithm converged in 36 iterations (about $34s$ of CPU time in MATLAB on a Pentium II at 233 MHz). A classical ANOVA takes a little over $0.2s$ in the same environment.

To check our model we plotted both the well effects and the residual errors versus concentration (not shown). The well effects show no evidence of correlation with concentration. There is evidence that the residuals tend to be

[5] The numerical values of the concentrations are not evenly spaced; instead, they were chosen by biologists to cover a wide range of effects.

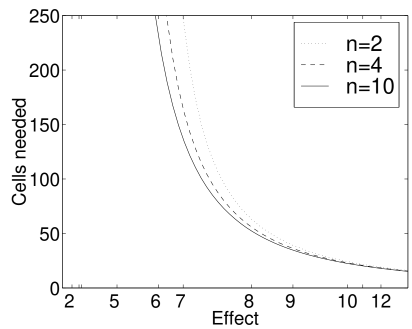

Figure 4: Cells needed to detect differences from lowest concentration. Horizontal tics are estimated responses 2 through 12 from Figure 3.

slightly larger at higher concentrations, but not enough larger to cause us to worry about the validity of our conclusions.

10.2 Calculations

Now that we have the fitted model, we can calculate the answer to our question about how to design future experiments to detect the drug.

We will accept 5% false positives and 5% false negatives in our future experiments. Each experiment will measure $n/2$ treatment and $n/2$ control wells, with N/n cells per well. We will say that the experiment detects the drug if the measured difference between treatment and control wells exceeds some threshold.

Call the expected difference between treatment and control Δ and the observed average difference D. The variance of D is $S^2 = 4(\sigma^2/N + \alpha^2/n)$. To see why, write y_i for the output of the ith cell in the new experiment, ϵ_i for $y_i - E(y_i|\text{well, treatment})$, w_j for the effect of the jth well in the new experiment, and (i) for the index of the well containing the ith cell. Now we can write D as

$$\begin{aligned}
&\frac{2}{N}\left(\sum_{\text{treatment}} y_i - \sum_{\text{control}} y_i\right) \\
&= \frac{2}{N}\sum_{\text{treatment}}(\epsilon_i + w_{(i)} + \Delta) - \frac{2}{N}\sum_{\text{control}}(\epsilon_i + w_{(i)}) \\
&= \frac{2}{N}\sum_{\text{treatment}}\epsilon_i - \frac{2}{N}\sum_{\text{control}}\epsilon_i + \frac{2}{n}\sum_{\text{treatment wells}} w_j \\
&\quad - \frac{2}{n}\sum_{\text{control wells}} w_j + \Delta
\end{aligned}$$

The first four terms are independent, while the last is deterministic. The vari-

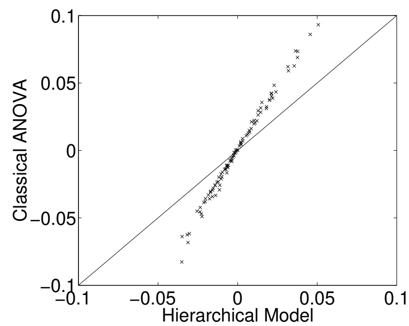

Figure 5: Well effects from the hierarchical model vs. those from the classical ANOVA. Solid line is $y = x$.

ance of each of the first two terms is $2\sigma^2/N$, while the variance of each of the next two is $2\alpha^2/n$. Adding these variances together gives the desired result.

To avoid false positives, the threshold must be bigger than $1.96S$. To avoid false negatives, Δ must be an additional $1.96S$ beyond that. So, we need $\Delta \geq 7.84\sqrt{\sigma^2/N + \alpha^2/n}$. Fixing Δ and n and solving for N gives the detection curves shown in Figure 4. The horizontal axis shows how big an effect we are trying to detect; the vertical axis is the number of cells we need to measure; and the three curves correspond to splitting the cells into 2, 4, or 10 wells.

11 Comparison with ANOVA

For comparison, we ran a classical ANOVA as described in Sections 3 and 4. The classical ANOVA doesn't handle our singular design matrix very well, so we tried two different fixes, both of which gave essentially the same result. The first fix is to place a weak prior on the well effects; this prior approximately picks, from the infinitely many w vectors which maximize the likelihood, the one with the smallest weights on the well features. (The concentration and plate effects will move to accommodate this requirement.) Then we can use a still weaker prior on the concentration effects to pick out the w vector with the smallest weights on the concentration features. (The plate effects will move to accommodate this requirement, while the well effects will approximately stay put because their prior is stronger.) This fix is inelegant since it is sensitive to the relative strengths of the weak priors.

The second fix is to add a strong prior that the well effects will sum to nearly zero within each concentration and within each plate. To construct such a prior, let H be the matrix whose columns are the 14 constraints on w. For example, one column of H will have 1s in the rows corresponding to the wells exposed to dose number 7 and 0s elsewhere. Then the appropriate prior is that w is normally distributed with mean zero and precision kHH^{T} for some large k.

This prior is improper, but the data will provide information in every direction that the prior doesn't. This fix is inelegant since we have to select the scales of the columns of H arbitrarily.

Another possible fix (which we didn't try) would be to redefine X so that it has 14 fewer columns. This fix is also inelegant, since it requires us to make arbitrary choices about which linear combinations of features to keep.

Both of the above variants of classical ANOVA return essentially the same answer, although they differ in the tradeoff between concentration effects and the constant term. Their estimate of σ is nearly the same as the EM algorithm's, as is their estimate of the concentration effects after correcting for the constant term.

More importantly, the classical ANOVA differs from EM on the estimates of the well effects and their variance. The ANOVA's well effects are larger than those calculated by EM, as shown in Figure 5. This happens because the EM algorithm puts a prior on the well effects to compensate for the lack of data in estimating them. And, the classical ANOVA's estimate of α is too small: it puts α at 0.0252. This value is about 8% smaller than the one computed by EM.

12 Discussion

We have described the theory and computation necessary to fit a Bayesian hierarchical linear model, and we have demonstrated the application of this model to real-world biochemical data. Hierarchical models have several advantages which ordinary Bayesian linear regression lacks, including the ability to ask questions about groups of weights, the ability to specify that several weights are likely to be similar in size without limiting that size, and the ability to model experiments where the number of regressors grows with the number of samples. We took advantage of all three of these properties while analyzing our biochemical data.

Acknowledgements

Thanks to Andrew Moore both for helping me get access to the data analyzed in this paper and for helpful comments on drafts. This work was supported by NSF KDI award number DMS–9873442. The opinions and conclusions are the author's and do not reflect those of the US government or its agencies.

References

[1] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, 34:1–41, 1972. Includes discussion.

[2] Stephen F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 53–72.

Kluwer Academic, 1989. Proceedings of the Eighth MaxEnt Workshop, Cambridge.

[3] David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[4] Thomas P. Minka. Bayesian linear regression. Unpublished manuscript, available from `http://www.media.mit.edu/~tpminka`, 1999.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, London, 1995.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–37, 1977.