Hierarchical Models

Statistics 220

Spring 2005

Hierarchical Models

Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters

1. Data model: $y|x, \theta_y$

2. Process model: $x|\theta_x$

3. Parameter model: θ_x, θ_y

Note 1: actually all of the models we have seen so far have been hierarchical, but most only had two levels to the hierarchy.

Note 2: there may be a hierarchical structure within each piece. For example, the process model may involve a time series model. So the full model may involve more than three levels in the hierarchy.

Note 3: A three level hierarchical model doesn't have to fit this structure. For example

$$\begin{aligned}
y_i|\mu,\sigma^2 &\overset{iid}{\sim} N(\mu,\sigma^2) \quad i=1,\ldots,n \\
\mu|\sigma^2 &\sim N\left(\mu_0,\frac{\sigma^2}{\kappa_0}\right) \\
\sigma^2 &\sim \mathrm{Inv}{-}\chi^2(\nu_0,\sigma_0^2)
\end{aligned}$$

This three level approach is a common way of presenting hierarchical models (Berliner, 2000; Clark, 2005 [on web site]; etc)

Why go hierarchical?

- Non-hierarchical models with few parameters generally don't fit the data well.

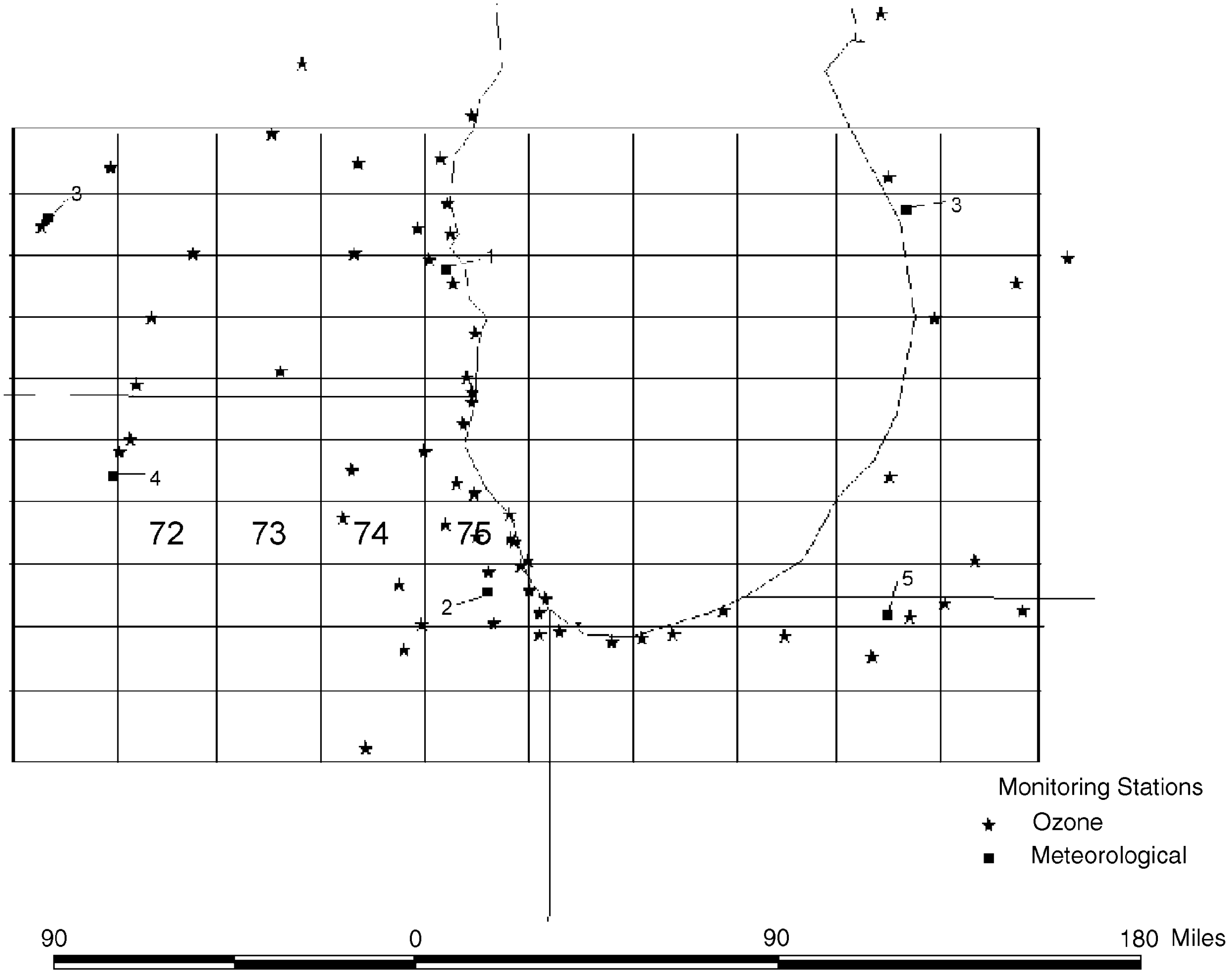
- Non-hierarchical models with many parameters then to fit the data well, but have poor predictive ability (overfitting)

- Hierarchical models can often fit data with a small number of parameters but can also do well in prediction.

- Hierarchical models with more parameters than data points can be useful and can give reasonable answers

An example of a hierarchical model is given in McMillan et al (2005) (Available on the articles page). It describes a predictive model for daily ground level ozone given meteorology in the Lake Michigan region of the Midwest.

1. Data: $Z_t =$ maximum of 8 hour average ozone measured at 58 stations, $M =$ meteorology at 6 locations (daily data).



2. Process: O_t = true mean maximum 8-hour average ozone over a 10 × 10 grid (spatial mean taken over grid box).

3. Parameters: measurement error variance (σ_z^2), meteorology regression parameters (β), regression parameters relating true ozone on day t to true ozone in neighbouring grid boxes on day $t-1$ (θ), time varying mean ozone intercept (μ_t), ozone process variance (σ_o^2)

The form of the model is

1. Data model: $Z_t = KO_t + N(0, \sigma_z^2 I)$ where K is a mapping matrix which indicates which grid box a measurement was made.

2. Process model:

$$p(O_{1:T}|M_{1:T}, \boldsymbol{\theta}_o) = p(O_0|M_0, \boldsymbol{\theta}_o) \prod_{t=1}^{T} p(O_t|O_{t-1}, M_{1:t}, \boldsymbol{\theta}_o)$$

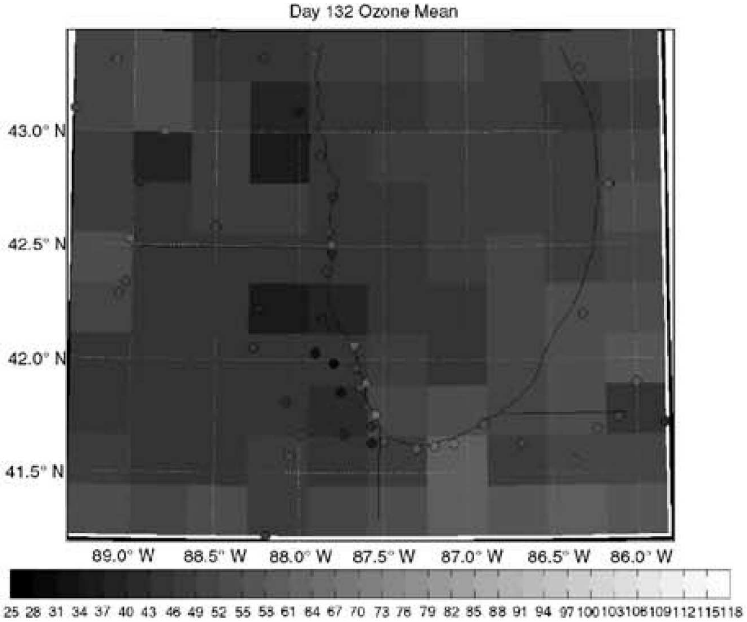
where

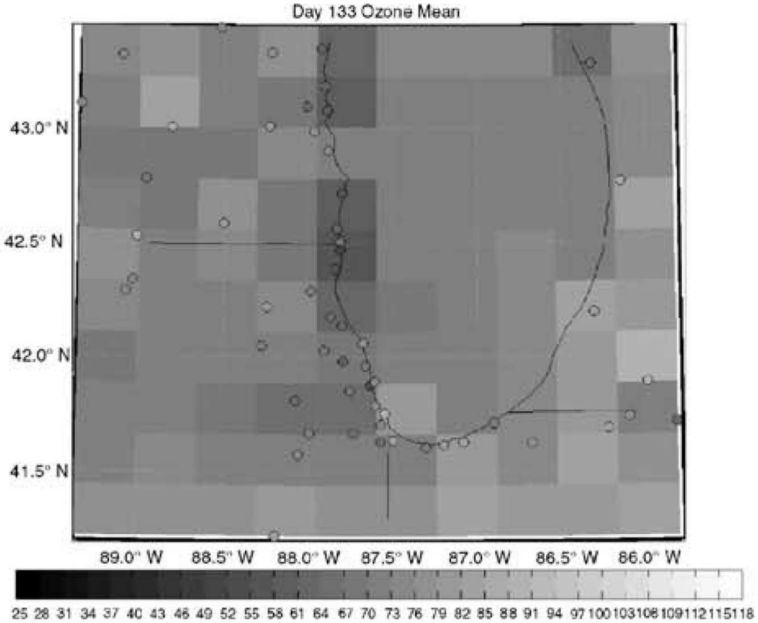
$$p(O_t|O_{t-1}, M_{1:t}, \boldsymbol{\theta}_o) = \mu_t \mathbf{1} + H(\theta)O_{t-1} + G(\theta)B_{t-1} + M_t\beta + N(0, \sigma_o^2 I)$$

μ_t is modeled as two-regime process: "normal" or typical behaviour and "high pressure" system behaviour. Conditional on which regime is active, μ_t is modeled as a first-order, autoregressive time series with regime-dependent mean and autoregression parameters. Regime states and transitions are then modeled as a first order Markov chain, whose transition probabilities depend on a recursively filtered, areally averaged air pressure series.

3. Parameter model:

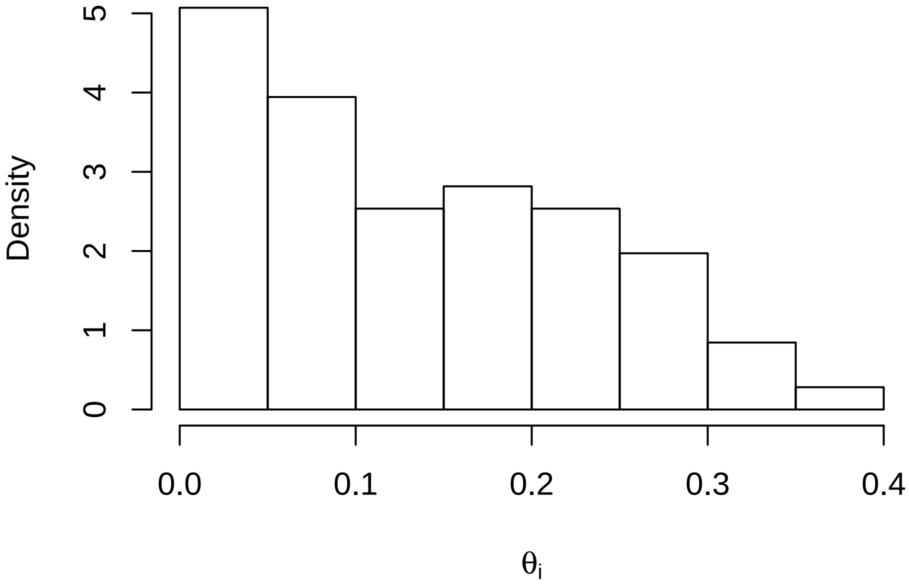
   Generally vague priors were put on the parameter values, except for parameters for which vague prior lead to poor posteriors. This problem occurred with the $H(\theta)O_{t-1}$ terms. Also it was necessary to designate one regime as "normal" and one as "high pressure" through average ozone observation (needed to specify regime for first day).


Day 132 Ozone Mean


Day 133 Ozone Mean

Example: Tumor rates in rats

71 different groups rats. Interest in the rate of endometrial stromal polyps in the different groups. The number of rats varies from group to group.



- Data model: y_i = number of tumors in group i

$$y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \ldots, 71$$

- Process model: θ_i = tumor rate in group i

$$\theta_i \overset{ind}{\sim} Beta(\alpha, \beta)$$

- Parameter model: $\alpha, \beta \sim p(\alpha, \beta)$.

The idea behind this model, is that we expect similarity behind the tumor rates in the different groups, but we don't expect them to be exactly the same. For example, the experimental conditions won't be exactly the same (e.g. different batches of rat chow fed to the different groups, drift over time, etc)

With a model like this, we can "borrow strength" from the other groups to come up with better estimates for each of the θ_i.

If we know α and β, we would expect the θ_i's to cluster around $\frac{\alpha}{\alpha+\beta}$. There will be some variation about this (prior variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$). But for a particular i, how far θ_i might vary from $\frac{\alpha}{\alpha+\beta}$ will be influenced by y_i and n_i.

However we don't know α and β. However given the data we can estimate them.

Setting up Hierarchical Models

While the Data, Process, Parameter framework is useful, it doesn't tell you what distributions to plug into a different problem.

Some parts will be "obvious", such as $y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta)$ in the rat tumor example.

However, other parts won't be. For example, in the ozone example is

$$Z_{it}|O_t, \sigma_z^2 \overset{ind}{\sim} K_i O_t + N(0, \sigma_z^2)$$

reasonable.

Two possible questions may come to mind here

1. Is the independence of the measurements errors, given the truth reasonable? Independence between the different ozone stations probably ok. Independence between days for the same station more questionable. Maybe some correlation between days might be more reasonable. For many problems, independence of measurement errors is a reasonable working hypothesis.

2. Is mean zero assumption of the measurement errors reasonable? Somewhat questionable. No reason to assume a constant ozone surface across a whole grid box. However if the grid boxes are small enough, this is probably a reasonable working assumption.

Exchangeability

A useful assumption in building models, if no information, other than the data y is available to distinguish any of the θ_j's from any of the others, and no ordering of grouping of the parameters can be made, one must assume symmetry among the parameters in the prior.

For example, in the rat tumor example, we have no prior reason to assume that $\theta_{70} < \theta_{71}$ is more likely than $\theta_{70} > \theta_{71}$. In fact, for the information given, the order that the groups are listed in is meaningless.

So for this problem, it seems reasonable to have the distribution on the θ_j's be exchangeable, i.e. the distribution $p(\theta_1, \ldots, \theta_J)$ should be invariant under permutations of the indices $(1, \ldots, J)$. If $J = 3$, then the distributions

$$p(\theta_1, \theta_2, \theta_3), p(\theta_1, \theta_3, \theta_2), p(\theta_2, \theta_1, \theta_3), p(\theta_2, \theta_3, \theta_1), p(\theta_3, \theta_1, \theta_2), p(\theta_3, \theta_2, \theta_1)$$

are all of the same form.

If there is information in the indices about the distributions, exchangeability is usually not reasonable. Suppose that different pure-bred rat strains were used for groups 50 to 71 than those used for groups 1 to 49. Then exchanging indices 49 and 50 would not be reasonable (probably).

Note that exchangeability does not imply independence. For example, the multivariate normal model

$$y \sim N_d(\mu\mathbf{1}, \Sigma)$$

where $\mathrm{Var}(y_j) = \sigma^2$ for all i and $\mathrm{Corr}(y_i, y_j) = \rho \neq 0$ for all i and j, is exchangeable, but obviously not independent.

Exchangeability implies the marginal distributions for each component are the same (identically distributed), but nothing about independence. In fact the dependence between the different components must be the same.

However all iid models are exchangeable.

One way of getting exchangeable distribution is to take a mixture of iid distributions.

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

As ϕ is usually unknown, the distribution on θ must average over the uncertainty in ϕ.

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi) d\phi$$

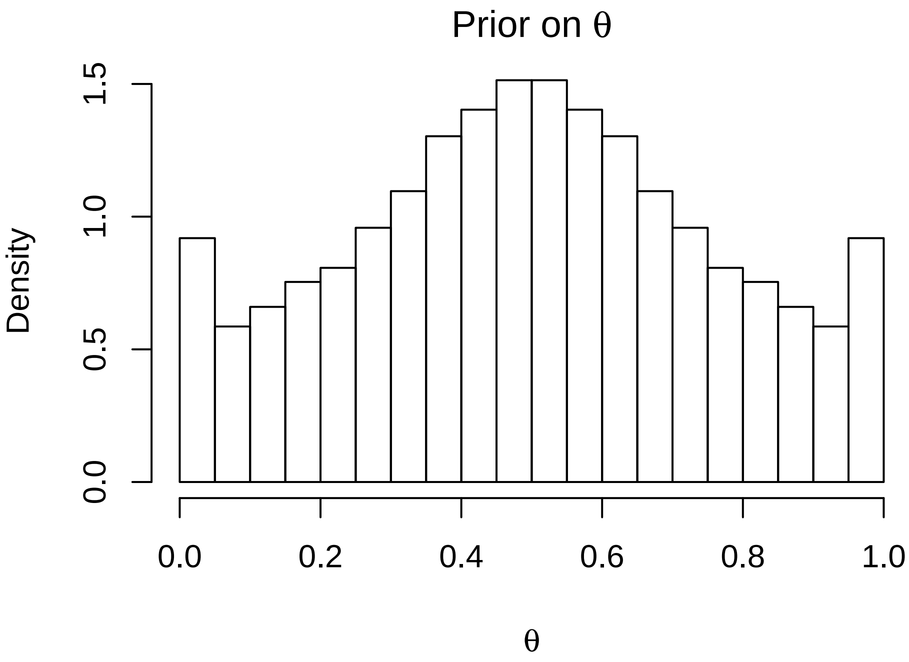
All models of this form are exchangeable. However for finite J, not all exchangeable models can be written in this form. de Finetti's theorem states as $J \rightarrow \infty$ any well behaved exchangeable distribution can be written in this form.

One way to think of this approach to get an exchangeable model is to think of the θ_j's as draws from a superpopulation model that is determined by the hyperparameter ϕ.

One way of thinking of exchangeability is in terms non-informativeness or ignorance about the random variables. In the rat example, we have no preferences for different orderings of the theta's. Note these distributions are not non-informative. For example in the rat problem, a model like

$$\begin{aligned}\theta_i &\overset{iid}{\sim} Beta(\alpha, \beta) \\ \alpha &\sim U(0, 20) \\ \beta &\sim U(0, 20)\end{aligned}$$

is highly informative for θ.



Inference in Bayesian Model

Suppose we have the following hierarchical model

$$\begin{aligned} y|\theta,\phi &\sim p(y|\theta) \\ \theta|\phi &\sim p(\theta|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

The joint prior is

$$p(\theta,\phi) = p(\phi)p(\theta|\phi)$$

and the joint posterior is

$$\begin{aligned} p(\theta,\phi|y) &\propto p(\phi)p(\theta|\phi)p(y|\theta,\phi) \\ &= p(\phi)p(\theta|\phi)p(y|\theta) \end{aligned}$$

Choice of the hyperprior on ϕ needs some thought. Often it will be a non-informative prior. However care needs to be taken, as a improper prior can lead to a improper posterior. For example, in the ozone example, all priors are proper. The problem here is that there are more parameters than data points so an improper prior may lead to problems in this example. (Not explicitly checked, but the results of the MCMC chain suggests problems)

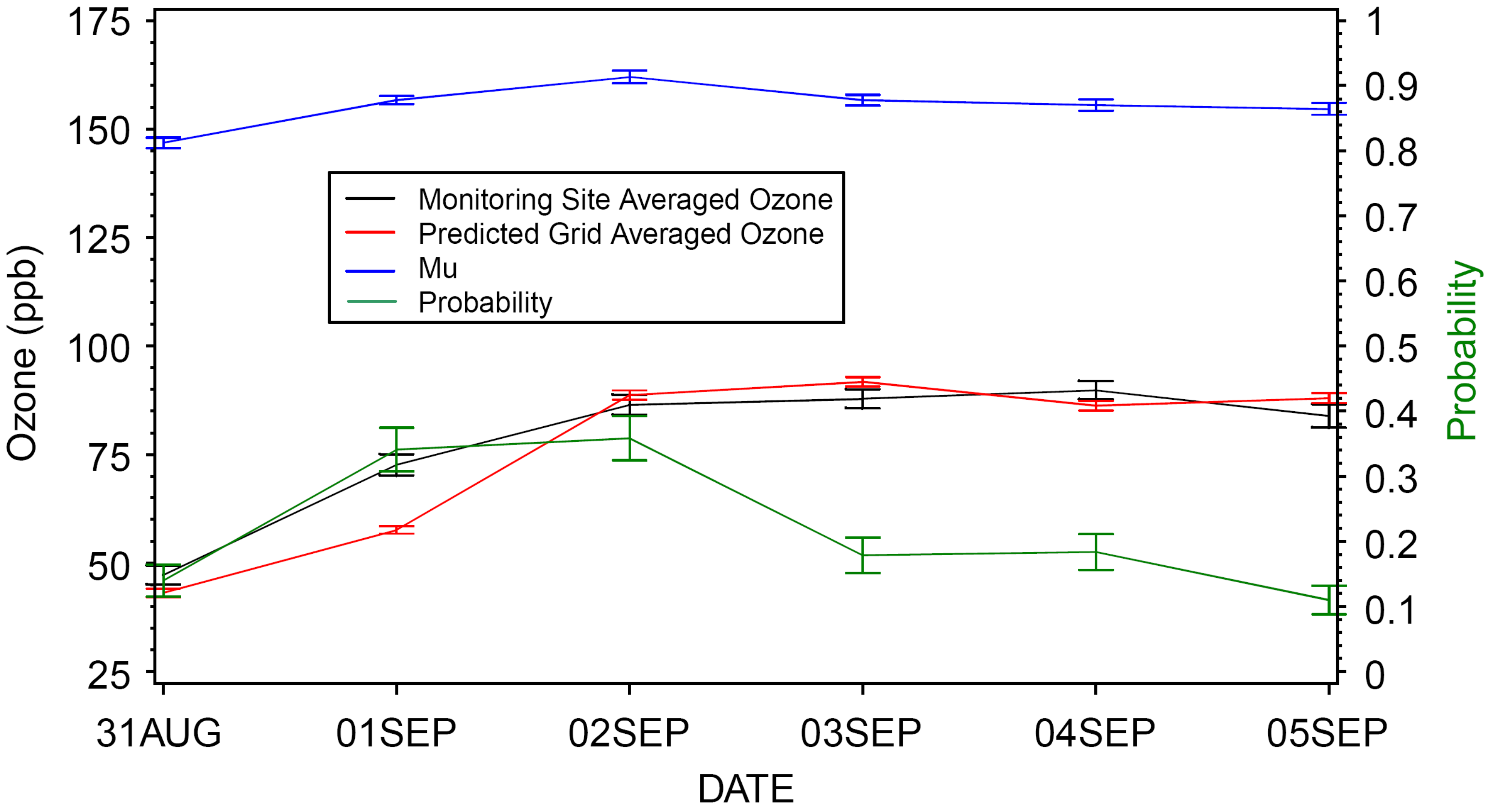
Of interest in a hierarchical model are posterior predictive distributions. There are two situations of interest

1. $\tilde{y}$ for an existing θ_j

2. $\tilde{y}$ for a new θ_j

The first situation could occur in the rat example. Suppose we had an addition n rats from group j. Then we would expect $\tilde{y}|\theta_j \sim Bin(n, \theta_j)$.

An example of the second situation is the ozone model. The purpose of this model was to develop a predictive model for ozone on day $t+k$ given the ozone and meteorology data from days 1 to t and meteorology predictions for days $t+1$ to $t+k$.


Monitoring Site Averaged Ozone
Predicted Grid Averaged Ozone
Mu
Probability
Ozone (ppb)
Probability
31AUG
01SEP
02SEP
03SEP
04SEP
05SEP
DATE