5 Ensemble QPF - A Statistical Model

These notes and other files and links can be found at http://www.meteor.iastate.edu/~jonhobbs/R/.

This session brings together some of the ideas from previous sessions to develop and assess a statistical model for a problem of interest. As I see it, some key aspects of this process include

- Initial Considerations
- Exploratory Analysis
- Formulating the Model
- Estimating the Model
- Model Assessment

5.1 The Problem

Ensemble forecasts provide an opportunity for probabilistic forecasting. One variable of key interest is precipitation. However, ensembles can be under-dispersive and/or biased, as exhibited by our rank histogram example from a previous session. These problems directly impact any probabilistic statements made about the ensemble.

![Rank Histogram](image)

One possibility for addressing these problems is to formulate a statistical model to describe the situation and to account for some of the inherent variability. Ultimately, we would like to describe the predictive distribution of the observed precipitation. The predictive distribution is the probability distribution (i.e. pdf) of the quantity of interest (observed precipitation) given some predictor(s), which would be one or more ensemble-member forecasts.

Mathematically, we could write the predictive distribution as

$$ f(y_i|x_i, \theta_i). $$

Here, $y_i$ is the observed precipitation for location $i$. The ensemble forecast(s) are represented by the vector $x_i$ and the vector $\theta_i$ is a set of parameters that represent the predictive distribution (more on this later).
5.2 The Data

The dataset we will be examining is a set of 24-48 hour precipitation forecasts from July 27-August 15, 2005 and the verifying Stage IV analyses. The forecasts come from a six-member “poor-man’s ensemble” (Ebert, 2001) constructed from several mesoscale models. For the 20-day period, the 00Z model runs were used, and all forecasts and Stage IV data were re-gridded to a common grid using the neighbor-budget procedure (Accadia et al., 2003).

Sources for the “poor-man’s ensemble”

<table>
<thead>
<tr>
<th>Model</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRF BMJ</td>
<td>ISU (A. Clark)</td>
</tr>
<tr>
<td>WRF KF</td>
<td>ISU (A. Clark)</td>
</tr>
<tr>
<td>Eta BMJ</td>
<td>ISU (A. Clark)</td>
</tr>
<tr>
<td>Eta KF</td>
<td>ISU (A. Clark)</td>
</tr>
<tr>
<td>MM5</td>
<td>ISU (D. Herzmann, D. Flory)</td>
</tr>
<tr>
<td>NAM</td>
<td>NCEP</td>
</tr>
</tbody>
</table>

Aside: Each of these models (and Stage IV) produces some sort of binary output files (Grib or NetCDF for the most part), so there is some effort necessary to make everything “R-ready.” My typical method of choice is to have a single NetCDF file for each day as the final product.

5.2.1 The Goal

The goal for this dataset is to describe the predictive distribution adequately for the 20 days of data available and then use this description to predict the 21st day based on the ensemble forecasts for that day. Of course the 20-day window is arbitrary here, and ideally one would want to devise a method to find the optimal “training period.”

5.3 Preliminary Considerations

Before looking at the data, it might be useful to return to the formulation for the predictive distribution \( f(y_i|x_i, \theta_i) \) and answer some preliminary questions. Some questions that come to my mind are

1. What are some possible values for \( y_i \)?
   - We are dealing with precipitation over a 24-hour period. We should expect some zeros and some positive values. Clearly, negative observations are not possible.
   - Regardless of the values of \( x_i \), there is probably some chance that \( y_i \) will be zero. This leads to the next question.

2. As the values of \( x_i \) change, how does \( f(y_i|x_i, \theta_i) \) change?
   Here, it is probably useful to consider just a single value of \( x_i \), not the entire vector of forecasts \( x_i \).
   - I would expect \( P(y_i = 0) \) to decrease as \( x_i \) increases.
   - I would expect that the mean of \( f(y_i|x_i, \theta_i) \) would increase as \( x_i \) increases. If not, we are in some serious trouble!

5.4 Exploratory Analysis

A bit of philosophical discussion will start this section. I take “exploratory analysis” to mean an initial look at the data through graphical methods and perhaps some basic statistics, such as means, variances and frequencies. Exploratory analysis can (and probably should) be a fairly in-depth and time-consuming process for complicated or large datasets. I believe that this is a very important part of the statistical modeling process; it does not constitute the whole process, but it should not be overlooked either.

Philosophy aside, what can we accomplish through exploratory analysis? I like to sum this up with one broad question, which will be phrased for our example:

- Empirically, what does \( f(y_i|x_i, \theta_i) \) look like?
It is important to keep in mind that the predictive distribution is a *conditional distribution*, meaning that it is specified in terms of other quantities, most importantly the $x_i$. That suggests that any exploratory analyses of $y_i$ should be for specific values or ranges of $x_i$.

One way to completely display a conditional distribution is with a scatterplot. Here we display some scatterplots for different ensemble members (supplement).

```r
plot(fcsts$etakf,fcsts$st4,main="Eta KF",xlab="Eta KF",ylab="Stage IV",
pch=16,cex=0.4,col="blue")
```

These scatterplots do not seem particularly informative, although they do seem to confirm both points under question 2 above.

For further investigation, we can use the subsetting capabilities of R to investigate the observations $y_i$ for specific values of $x_i$. Histograms, frequencies, means and variances for these subsets all can prove useful.

After examining some of this exploratory analysis, it seems that it may be useful to view the predictive distribution as a combination of two different components.

1. There is a positive probability of observing precipitation, $P(y_i > 0)$ that increases non-linearly with the value of the forecast.

2. On average, the nonzero precipitation amounts increase as the value of the forecast increases. The variance of the observed precipitation increases with the mean.

### 5.5 Formulating a Model

Now, we will attempt a mathematical formulation of the predictive distribution based on the results of the exploratory analysis.

Precipitation amounts can be modeled as a mixture with two components. The “mixing” distribution is for a binary rain/no rain event and the data model is a distribution for the actual precipitation amount. Both are conditional on the forecast.

1. **Mixing distribution**
   - $R$ is a binary random variable defined as
     
     $$ R = \begin{cases} 
     0 & \text{if no rain observed} \\
     1 & \text{if rain observed} 
     \end{cases} $$

   and the distribution for $R$ is
     $$ g(r|p) = p^r (1 - p)^{1-r} $$

   The observed binary event can be related to a forecast value $x_i$ through a logistic regression.
     $$ g(r_i|\lambda_i) = \exp [r_i\lambda_i - \log\{1 + \exp(\lambda_i)\}] $$

   $$ \lambda_i = \alpha_0 + \alpha_1 \sqrt{x_i} $$

   The square root transformation was arrived at empirically, based on examination of observed data (see the “Probability of Precipitation” plots).

2. **Data distribution**
   - The distributional form for precipitation amount $Y$ is different, depending on the value of $R$:
     
     $$ f(y|r, \theta) = \begin{cases} 
     I(Y = 0) & \text{if } R = 0 \\
     c(y|\theta) & \text{if } R = 1 
     \end{cases} $$

   The function $c(y|\theta)$ is some continuous distribution on $(0, \infty)$. Here it will be taken as a gamma distribution, from which a generalized linear model can be formulated. Specifically,

   $$ c(y_i|\mu_i, \phi) = \exp \left[ \phi \left\{ -\frac{1}{\mu_i} y_i - \log(\mu_i) \right\} \right. $$

   $$ + \phi \log(y_i) - \log(y_i) + \phi \log(\phi) - \log(\Gamma(\phi)) $$

   $$ \mu_i = \beta_0 + \beta_1 x_i $$

3
The likelihood then becomes

\[ L(\alpha, \beta, \phi|x, y) = h(y|x, \theta, p) = f(y|x, \theta)g(r|p), \]

and after some manipulation, the log likelihood (from which maximum likelihood estimates can be found) is

\[ \ell(\alpha, \beta, \phi) = \sum_{i=1}^{n} \log \{g(r_i|\lambda_i)\} + \sum_{r=1} \log \{c(y_i|\mu_i, \phi)\} \]

Since we are treating all observations as independent, the combined data log likelihood (for all observations) is simply the sum of the individual data log likelihoods.

### 5.6 Fitting the Model

Next time . . .

### References


# Read the various netcdf files

```r
library(ncdf)

c1=open.ncdf("st4comp.nc")
st4=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
st4=st4[24:58,14:33,]

c1=open.ncdf("wrfbmjcomp.nc")
wrfbmj=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
wrfbmj=wrfbmj[24:58,14:33,]

c1=open.ncdf("wrfkfcomp.nc")
wrfkf=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
wrfkf=wrfkf=wrfkf[24:58,14:33,]

c1=open.ncdf("etabmjcomp.nc")
etabmj=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
etabmj=etabmj[24:58,14:33,]

c1=open.ncdf("etakfcomp.nc")
etakf=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
etakf=etakf[24:58,14:33,]

c1=open.ncdf("mm5comp.nc")
mm5=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
mm5=mm5[24:58,14:33,]

c1=open.ncdf("namcomp.nc")
nam=get.var.ncdf(nc1,nc1$var[[1]])
close.ncdf(nc1)
nam=nam[24:58,14:33,]

fcsts = data.frame(cbind(st4,wrfbmj,wrfkf,etabmj,etakf,mm5,nam))
```
categories.R

gp = rep(0,length(fcsts$etakf))
cats = c(0,0.00001,0.5,1,1.5,2,5,7.5,10,20,Inf)
ctr = c(0,0.25,0.75,1.25,1.75,3.5,6.25,8.75,15,30)
 probgrp = function(c) {
   p = length(fcsts$st4[grp == c & fcsts$st4 > 0.00001]) / length(fcsts$st4[grp == c])
}

 meangrp = function(c) {
   m = mean(fcsts$st4[grp == c & fcsts$st4 > 0.00001])
}

 vargrp = function(c) {
   v = var(fcsts$st4[grp == c & fcsts$st4 > 0.00001])
}

 for (i in 2:length(cats)) {
   grp[fcsts$wrfbmj >= cats[i-1] & fcsts$wrfbmj < cats[i]] = i-1
 }

 ps = sapply(1:10,probgrp)
 ms = sapply(1:10,meangrp)
 vs = sapply(1:10,vargrp)

 for (i in 2:length(cats)) {
   grp[fcsts$wrfkf >= cats[i-1] & fcsts$wrfkf < cats[i]] = i-1
 }

 ps = c(ps,sapply(1:10,probgrp))
 ms = c(ms,sapply(1:10,meangrp))
 vs = c(vs,sapply(1:10,vargrp))

 for (i in 2:length(cats)) {
   grp[fcsts$etabmj >= cats[i-1] & fcsts$etabmj < cats[i]] = i-1
 }

 ps = c(ps,sapply(1:10,probgrp))
 ms = c(ms,sapply(1:10,meangrp))
 vs = c(vs,sapply(1:10,vargrp))

 for (i in 2:length(cats)) {
   grp[fcsts$etakf >= cats[i-1] & fcsts$etakf < cats[i]] = i-1
 }

 ps = c(ps,sapply(1:10,probgrp))
 ms = c(ms,sapply(1:10,meangrp))
 vs = c(vs,sapply(1:10,vargrp))

 for (i in 2:length(cats)) {
   grp[fcsts$mm5 >= cats[i-1] & fcsts$mm5 < cats[i]] = i-1
 }

 ps = c(ps,sapply(1:10,probgrp))
 ms = c(ms,sapply(1:10,meangrp))
 vs = c(vs,sapply(1:10,vargrp))

 for (i in 2:length(cats)) {
   grp[fcsts$nam >= cats[i-1] & fcsts$nam < cats[i]] = i-1
 }

 ps = c(ps,sapply(1:10,probgrp))
 ms = c(ms,sapply(1:10,meangrp))
 vs = c(vs,sapply(1:10,vargrp))