Tao Huang

Department of Mathematical Sciences
Montana State University

May 3, 2013

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Tao Huang

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	YOUR ADVISOR'S NAME HERE Writing Project Advisor
Date	Megan D. Higgs Writing Projects Coordinator

[journal=jacsat, manuscript=article]achemso
[version=3]mhchem
tao.huang@msu.montana.edu
[Montana State University] Department of Mathmatics
`achemso` demonstration] **An Alternative of Modelling in Spatial Statistics**

Abstract

It's a standard procedure to model statistical data using given variables and check residuals after fitting a model. However, according to the well established statistical theory, probability is a valid summary of some random occurrences of a certain event. In Cox's theorem, probability is taken as a primitive (that is, not further analyzed) and the emphasis is on constructing a consistent assignment of probability values to propositions.[1] In other words, if the assigned probability is valid, it's the same probability assigned to the occurrences of the same event, as it's defined as a consistent assignment. On the other hand, if the underlining probability of occurrences of the event changed, to assign a probability becomes questionable. Instead of trying to model, an alternative is proposed where we try to determine the additive components of probability based on grouping random responses and put the components together on a spatial grid to help understand the relationship between the set of explanatory variables and counts data. In the light o Cox's theorem, such a procedure focuses on consistent probabilities assigned first, based on which we model the data using existing variables to find hidden patterns.

1 Introduction

Consider probability theory in the context of flipping a coin which is a simplest case to demonstrate probability theory. For example, after 20 flips, we have the following sequence of tails and heads: 0 0 0 0 0 1 0 1 1 1 1 0 0 0 1 0 1 1 1 0. Given such a sequence, we are not confident at all to predict the outcome of a single flip. (In terms of the resolution later on introduced, the resolution is not high enough to see a single flip clearly.) To avoid such an embarrassment, we calculate the proportion of heads or of tails which in fact is a probability summary assigned to flipping such a coin with the assumption that all flips are . Is a proportion a good enough summary? How about if I collect all zeros and set them in the first part of the sequence and leave all the ones in the last half of the sequence? In this case, the proportion which is no different than it is for the original sequence. As this sequence may not look like a the outcome of a random process, I don't think the sample proportion is a good statistical summary for such a sequence.

However, we may ask another practically useful question like this: is the coin different after, say, half of the sequence? As we may have enough observations to be split into two halves, we may calculate the sample proportions of the first half and of the second half and compare the two sample proportions in the context of statistics.

If the same question is asked over and over again, we'll finally find that there's a lower limit of the number of flips to which we can confidently assign a sample proportion as the probability of the involved occurrences. Though we can calculate a sample proportion only if there's at least one observation, it makes no sense to assign either 0 or 1 to such a flip.

We may start reflecting on such a process which is fundamental in statistics: the proportion of the whole sequence is bound to be blessed with a small estimated random error, but have we lost some information when the whole sequence was seen as a sequence where the probability of tails or heads occurrence is consistent across the board. How about if the probability changes systematically over the sequence? How much maximum information we may have of the sequence? It's a dilemma: a smaller standard error with a sample proportion masks more of a systematical pattern across the sequence, while on the other hand more information given to the pattern across the sequence leads to more loss in the information of the whole picture.

Let' now consider a spatial setting. If we want to model a spatial data set using a Poisson model, what does it mean for counts to be randomly distributed across an area? Informally, it's a distribution across an area without clustering, or other regular patterns. If we can find some areas with such a nice property, we can confidently assign a Poisson parameter as a summary statistic to each area. The areas are then changed to represent probability components on the map. We may repeat the process to separate all the probability components which can be pieced together to create the whole picture of the data set.

1.1 Assumptions

We assume that a Poisson model is sufficient to model a spatial counts data set if we believe that the counts are produced based on some probability process. This is plausible, because if the counts are proportional to the area raised to some power we may see more counts in a smaller area than in a bigger one. If we do see some patterns contradicting the assumption, therefore, it's plausible that the patterns are introduced by some other unknown factors confounding areas.

We also assume that the effects brought about by factors at play are continuous on the spatial grid. This is reasonable because all effects are continuous on a certain scale or the effects with a discrete variable are continuous over its range of effect in a spatial data setting. For example, if we see jumps in the explanatory variable level it's because the scale we use to measure the explanatory variable is too big to to reveal smaller details which are in the middle of the jumps. However, this is not a problem since there's always something too small to be detected precisely, we simply see all the discontinuous patterns as noise that can balanced out by averaging. we may take am example to illustrate what we can do about the small discontinuous noise. Imagine that noise is small random dots all over an area with different colors indicating different levels of some measurement. If the scale is magnitudes coarser than the size of the dots, we just see a single color nested in the grid instead of so many dots, as our visual system cannot detect something arbitrarily small. In such a light, if the number of patterns is big, the averaged noise induced by such patterns is supposed to be stable and relatively constant across the grid.

1.2 Procedures

- **Finding Contours**

  The first step in the proposed method is to find out the random components where the data counts are distributed according to a Poisson distribution with a certain parameter. Given a data set, we don't have any difficulty figuring out the total counts on the spatial grid and how much area is involved. Then a Poisson parameter is estimated for the whole data set. This is the start from which we try to find the contours on the spatial map to indicate all the random components. In order to find the probability components, two criteria are imposed in an algorithm. The contours are moved continuously to give the areas having best random patterns. And, when the contours are moved the Poisson parameters of the affected areas are estimated. The parameter estimates are supposed to be consistent with the summary parameter estimate: this is the second constraint. In other words, the algorithm tries to find boundaries of different areas while keeping the estimated Poisson parameter consistent with the Poisson parameters on a more general level on the spatial grid.

  Next, we put values of the explanatory variables on the spatial grid separately, and we use their levels as the responses based on which we make new contours on the grid.

[scale=.55]xy

[scale=.5]con1 [scale=.5]con2

[scale=.5]con3 [scale=.5]con4

[scale=.5]con5 [scale=.5]con6

[scale=1.2]con7

The following plot is the final contour plot with original counts overlaid on it:

[scale=1.2]ticket1

- **Finding a Good Match Between the Counts and the Explanatory Variables**

  We proceed to match the counts contour with each explanatory variable contour to see which of the variables gives the best spatial match. If none of them can achieve this goal, we may combine the variables to create a single variable as the response. We may be able to achieve another contour map which can be used to find a possible spatial match with the counts contour map.

1.3 Suggested Algorithms

- **Finding a Grid Fine Enough to Estimate Possible Contours**

  Given that determining whether the counts over an area is computationally intensive, the algorithm for the above purposes can be computationally expensive, as every step of changing the contour needs to be checked for randomness. If we have a map which is similar to our final contour map, it can be superimposed on a grid as the starting point to find the final contours. Computation involved could be much more effective and less expensive, as the superimposed map serves as a guide to find the final contours. For instance, we may define how fine the grid system is regarding the goal, say, some criterion like one count per block on average could be likely to work. After we have a grid, we may collect the blocks to achieve the best collection of them in terms of finding random components. Or, we may use a density contour map as the start which is much less computationally expensive. The discrete grid and the density contours superimposed can serve as a guide map.

- **Iterative Methods**

  We may start by finding an approximate inner most or outer most contour first and calculate the estimated Poisson parameters for the two components involved. Since there are always some some random errors associated with the estimated Poisson parameters and the contours, we may calculate possible confidence intervals for all of them to achieve some bounds within which we try to find the best match under the two proposed criteria. (To define what is best is one of the major tasks in of future works.) After the first step is

achieved, we proceed to find additional contours using the same technique. We stop the process once the partition indicates a random spatial distribution across the area.

1.4 Potential Advantages

- **Mathematical Consistency**

  As the proposed algorithm always keeps the parameters affected in a certain step consistent with the rest by respecting the total counts produced by the estimated parameters on all levels, each level of Poisson distribution is consistent with the more general levels of Poisson parameters. In contrast to the proposed advantage, the traditional way of modeling may not be able to give us a total count that is always approximately equal to the original total count, because if we model the data using the variables given, the integral of the explanatory variables over the domains may give us a sum significantly different from the ones produced by the most general model.

  A Poisson distribution is linear,[2]in the sense that we can add the parameters to achieve a new Poisson distribution. Therefore it's mathematically valid to lay the random components over a more general level. In mathematical terminology, the idea may be expressed as follows:

  If X falls into Poisson(λ_1), and Y falls into Poisson(λ_2),

  then X + Y falls into Poisson(λ_1+λ_2), where X and Y are random variables represent random counts in this context, and $\lambda_{1,2}$ are the Poisson parameters.

- **Causality**

  Even if there's no random assignment of the explanatory variables at all, we are confident to say that there's a strong association or possible causal relationship between the counts and some, maybe, combination of variables if the match is good enough. This also facilitates interpreting the analysis to the client or an audience, as the whole process is visible.

- **Prevention of Statistical Modelling Abuse**

  Given the contours resulting from applying the algorithm, we may be able to say that this contour map is telling us how much resolution the data set can give us, in terms of probability. Because if we try to zoom in further, the new areas can no longer produce random spatial distributions, and as a consequence, we are not justified to assign summary parameters to the areas! As assigning probabilities is not justified, statistical modeling is not justified hence. This means the limits of a probability model.

  In contrast, a traditional model may give us an estimated response even if a explanatory variable is given a slightly different value in the model. This is questionable in terms of resolution the data set can give. As a matter of fact, statisticians know that there's error with all estimates, but they don't know

what's the maximum amount of information the data set can give, in other words they don't know how much detail the data set can give.

In fact, if we are given a data set, we may be able to find out some association between the response and the explanatory variables in most cases. Because when we model a data set, we try many ways to find the strongest association between the response and explanatory variables. If we put such a way of modeling in the context of multiple comparison, we may immediately notice that it's possible to find some relationship anyway, after so many trials. But the matching of the spatial contours is graphical than numeric, this may suggest that it's hard to match up the two contours. However, if we are not able to find a good match no matter what combinations we tried, it's fair to say that the variables are not enough to determine such a pattern. But whether this suggested method works and how to define a match are potentially challenging.

- **No Worries about Residuals and Whether Including Variables**

  As the first goal of such a method is grouping homogeneous counts in areas, residuals are in fact in a good shape not only in modeling but also in probability theory. Instead of fitting a model by trying different combinations of variables, we simply match the contour maps to find association between the response and explanatory variables. We can exclude the variables that don't give a good match without considering modeling which is likely to give us an overfitted modeling by choosing the "best" model.

- **No Worries about Correlation Structures**

  As we overlay the response area contours, the final contour map which gives us the highest resolution in the probability framework automatically gives us a structure which is produced by the variables centering at different locations. Therefore, the "correlated observations" can be explained by a commonly shared factor at play. For example, we reasonably assume that the observations on the same individual are correlated because of some shared errors with the individual. However, the shared errors are produced by some shared factors in the same individual. In fact, by extending the reasoning like this, we look past the boundaries of some specific object to gain natural boundaries of factors determining our responses.

1.5 Future Works

- Defining "a match"

  This is a challenging task as this may affect the effectiveness of the suggest method a lot and a criterion of a match is difficult. Either a graphical criterion or a numeric one could apply. If a numeric criterion is applied, the suggested method could end up doing the same thing as the usual modeling methods do.

- Materializing the Algorithm

  Given the amount of computation that could be involved, it could be impossible to implement the ideas if the program is not optimized in terms of computation. And there are many different approaches to the implementation of the ideas in computing. To consider all the possible approaches and to see if there's any improvement by combining some of them take lots of time and effort.

References

[1] WIKIPIDIA *Cox's theorem*

[2] Schabenberger, Oliver/ Gotway, Carol A *Statistical Methods for Spatial Data Analysis, ISBN: 9781584883227*