2015

# Using a Spiral to Estimate Spatial Lattice Model Parameters

Geoffrey Glidewell
*University of Montana - Missoula*, geoffrey.glidewell@umontana.edu

# USING A SPIRAL TO ESTIMATE SPATIAL LATTICE MODEL PARAMETERS

By

GEOFFREY PAUL GLIDEWELL

Undergraduate Professional Paper
presented in partial fulfillment of the requirements
for the University Scholar distinction

Davidson Honors College
University of Montana
Missoula, MT

Official Graduation Date: May 2015

Approved by:

Jon Graham, Faculty Mentor
Mathematical Sciences Department

**ABSTRACT**

Glidewell, Geoffrey, B.A., May 2015 Mathematics

Using a Spiral to Estimate Spatial Lattice Model Parameters

Faculty Mentor: Jon Graham

This project explores the autologistic model for spatially correlated binary lattice data and uses a one-dimensional spiral to approximate two-dimensional data. An example of this type of data is the presence of disease in plants in a lattice framework. Each plant is labeled "diseased" or "non-diseased," where the presence of disease in one plant might increase, decrease or not affect the likelihood of disease in a neighboring plant. In order to fit an autologistic model to real data, the method of maximum likelihood would ideally be used to estimate the model parameters for the entire lattice. However, the model form involves an intractable normalizing constant preventing this method from being used directly. Although multiple methods have been developed to estimate the model parameters, most notably Markov Chain Monte Carlo (MCMC) maximum likelihood, these methods either rely on approximations of the normalizing constant or ignore the inherent spatial correlation. To calculate the constant directly, every possible lattice realization must be tabulated. However, for even a small lattice of size 20x20, this would mean $2^{400}$ different realizations, which is far too many for even a modern computer to compile. This normalizing constant can be computed in theory using two statistics computed from the data: S, the number of diseased sites, and N, the number of neighboring diseased sites from each realization. This project explored a method of generating all S and N combinations for a linearized subset of the two-dimensional lattice, allowing for calculation of the normalizing constant for the subset. For data on a spatial lattice, a spiral of locations can be extracted and an exact normalizing constant for the spiral calculated. Unfortunately one spiral uses only half of the data so must be combined with results from the remaining locations. Further investigation is being done to compare this method to known approximation methods in order to determine its viability.

Using a Spiral to Estimate Spatial Lattice Model Parameters

A binary spatial lattice is a two-dimensional matrix containing only 1's and 0's. The 1's can signify presence or "diseased" and 0's can signify absence or "non-diseased"; for example it can signal the presence/absence of a disease called phytophthora root and crown rot in bell peppers. The goal of the research project was to find an improved method of estimating the model parameters for the autologistic model used with binary spatial lattice data. Other estimation methods are currently in use, such as Markov Chain Monte Carlo (MCMC) maximum likelihood and pseudolikelihood, but the first method approximates the likelihood and the second ignores the spatial dependence which is likely present. It turns out that estimation of the model parameters relies entirely on the values of two statistics, S and N, that can be computed for a given set of binary lattice data, S is defined as the number of diseased sites in the lattice and N is defined as the number of neighboring pairs of diseased sites. They are sufficient statistics for this distribution, which means S and N contain the same amount of information about the model parameters as the full data. However, in order to use the S and N statistics to estimate model parameters, every possible (S,N) combination and its frequency of occurrence must be calculated because the normalizing constant for the autologistic model is a sum over all such combinations. Even a modern computer would not be able to do this for a reasonable sized 20x20 lattice, as it would require $2^{400} \approx 10^{120}$ calculations. The goal of this project was to find a manageable way to generate every possible (S, N) combination and its frequency of occurrence for any size lattice.

| | | |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 1 |

Figure 1: Example of 3x3 binary spatial lattice. For this lattice, the S statistic is 5, and the N statistic is 3 (two horizontal pairs and one vertical pair).

The autologistic model, shown in the figure below, gives the probability of any realization of the binary lattice data as a function of S and N, and as a function of the two model parameters α and β. The method of maximum likelihood finds those values of the parameters that maximize the probability of observing the actual data. The numerator of this autologistic form uses the S and N pair for the lattice data being analyzed, and the denominator is the normalizing constant. The normalizing constant is the sum of the given exponential function over every possible (S, N) pair and its

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}|\alpha,\beta) = \frac{e^{\alpha S+\beta N}}{\sum f_i e^{\alpha S_i+\beta N_i}}$$

Figure 2: The autologistic model. $\boldsymbol{Y}=\boldsymbol{y}$ is the given lattice of 1's and 0's. α and β are the model parameters. $f_i$ is the frequency of occurrence of the $i$th S,N pair.

frequency of occurrence. In order to perform maximum likelihood on the autologistic model, this normalizing constant must be known, which necessitates knowing the frequency of occurrence of every possible (S, N) pair for a lattice. This is the fundamental problem with maximum likelihood on this model. If the normalizing could be found, however, it would allow for the autologistic model parameters to be estimated in a straightforward manner.

The initial step was to write 3x3 matrices, starting with a matrix of only 0's and adding 1's sequentially, to eventually account for all $2^9$ possible 3x3 lattices. First, only the series of S values were generated. A relatively simple recursion was observed for computing S. The series of S values appeared as 0,1,1,2,1,2,2,3,1,2,2,3,2,3,3,4,1,2,2,3,… Although not obvious at first, this is a simple recursion pattern starting with a vector of length 1, the number 0. The recursion pattern is to first add 1 to the every value in the vector, then bind the new vector to the end of the initial vector. So 0 becomes 0,1. Repeat the recursion and 0,1 becomes 0,1,1,2. Repeat the recursion again and the vector becomes 0,1,1,2,1,2,2,3. Keep repeating until the vector is the length of the number of possible lattices (e.g. for a 3x3 this would be a vector of length $2^9$=512 values).

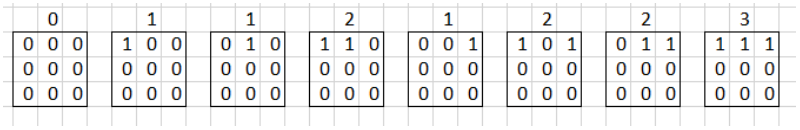


Figure 3. Example of sequential 3x3 matrices counted in order to generate a series of S values. The S value for each matrix is shown.

The process was repeated counting the N values instead of the S values. The series of N values also gives a sequence that can generated by recursion. However, as the N sequence is a function of the size of the lattice, unlike the S sequence, it requires two separate recursions. The first recursion generates an initial vector and the length of this vector is a function of the width of the lattice. Generation of the initial vector is started from the vector 0,0. For the *n*th recursion step, make a vector of $2^{n-1}$ 0's followed by $2^{n-1}$ 1's, add the vector to the original vector to generate a second vector, and bind the second vector to the initial vector. For example, 0,1 is added to 0,0 to give 0,1, then bound to return 0,0,0,1. Next 0,0,1,1 is added to this vector to give 0,0,1,2, which is bound to return 0,0,0,1,0,0,1,2. For a lattice of width *w*, this recursion step must be repeated *w* times.

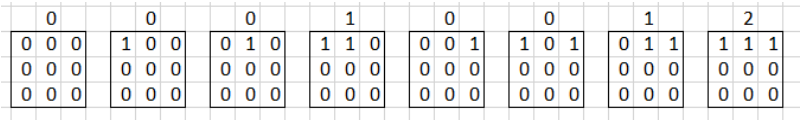


Figure 4. Same sequential 3x3 matrices shown above, but with N values instead of S values.

Once the initial vector is obtained, the second recursion must be used to generate the rest of the N values. For the *n*th recursion step, we alternate $2^{n-1}$ 0's and $2^{n-1}$ 1's for half the length of

the vector followed by alternating $2^{n-1}$ 1's and $2^{n-1}$ 2's. This recursion is a function of the initial width of the lattice, as for an *l* x *w* lattice every *w*th recursion must be a vector of only 0's and 1's, without the vector of 1's and 2's. So it would be alternate $2^{n-1}$ 0's and $2^{n-1}$ 1's for the full length of the vector.

Using these S and N vectors obtained by recursion, the normalizing constant can be calculated for a lattice by tabulating the S and N pairs and counting the frequency of occurrence of each pair. However, as the lattice size increases, the length of these vectors increases exponentially. The largest lattice for which a normalizing constant could be obtained using this method was a 5x5. Larger lattices quickly crashed the computer. But by generating tables of the frequency of occurrence of S and N pairs in smaller lattices, a pattern can be searched for that would hopefully extend to larger lattices.

Many tables of the frequency of occurrence of S and N pairs were generated and compared in search of a consistent pattern. Such a pattern was observed in lattices of width 1. It was noticed that every frequency number of an S and N pair from a lattice of width 1 was divisible by a binomial coefficient in Pascal's triangle, and these numbers appeared in a systematic fashion. The equation in Figure 5 is the equation eventually found to generate the frequency of any given (S, N) pair. By cycling through every possible (S, N) pair, the normalizing constant for the autologistic model can be quickly generated for a linear lattice of any length *l*.

$$freq = \binom{S-1}{N} * \binom{l-S+1}{S-N}$$

Figure 5. The equation used to find the frequency of occurence of every possible S, N pair in a lattice of wdith *l*.

As most real spatial data are two-dimensional, and linear lattices are one-dimensional, ideally a method could be found to treat spatial data linearly. The first method explored to do this was to "pull out" a spiral from those data, and estimate the parameters for this spiral using maximum likelihood. Although these parameters would best describe the data for the spiral under an assumed autologistic model, they would not be estimates of the parameters for the entire lattice. Ideally, a spiral could be found that had the same S statistic and the same N statistic as the entire lattice, but this is not possible, as the spiral can only account for one direction at a time (it is one-dimensional).

Pseudolikelihood is the most common method used currently to estimate parameters for the autologistic model. Pseudolikelihood assumes the binary responses at the sites are independent, and using the number of neighboring 1's for each site ($n_i$ is the number of neighboring 1's for site i in the equation in Figure 6), maximizes the product of the site-by-site probabilities of observing a 1 or 0 at the site given local spatial information. If the responses did not exhibit spatial correlation, this PL function is a true likelihood. Pseudolikelihood works well when the spatial dependence is not too large, but conceptually is not the right thing to do since the responses are inherently spatially dependent. Although using spirals would ignore many possible N pairs, it would not ignore the dependence between responses at neighboring sites.

$$PL(\alpha,\beta) = \prod_{i=1}^{m} \Pr(Y_i = 1|y_i) = \prod_{i=1}^{m} \frac{\exp\{\alpha + \beta n_i\}}{1 + \exp\{\alpha + \beta n_i\}}$$

Figure 6. The pseudolikelihood model for binary spatial lattices. $n_i$ is the number of neighboring 1's for the *i*th site in the lattice.

Initial analysis of previously analyzed data on Phytophthora root and crown rot (Gumpertz et al. 1997) returned parameters similar to those returned using pseudolikelihood. However, when analyzing lattices generated by the Gibbs sampler (a common technique for generating spatial lattices using preset parameters) a strong bias in the resulting parameter estimates was observed. When analyzing spirals instead of a full lattice, the parameter estimates returned were much lower than the parameter estimates obtained using pseudolikelihood. In order to accout for the fact that the spiral contains only half of the possible N pairs, the (S, N) sufficient statistics were adjusted in an effort to equate the information used in these two estimation processes. As indicated in Figure 7, the spiral considers many fewer possible N pairs. However, the bias still remained, returning parameters significantly different from the initial values.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 21 |
| 75 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 95 | 22 |
| 74 | 143 | 204 | 205 | 206 | 207 | 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 161 | 96 | 23 |
| 73 | 142 | 203 | 256 | 257 | 258 | 259 | 260 | 261 | 262 | 263 | 264 | 265 | 266 | 267 | 268 | 219 | 162 | 97 | 24 |
| 72 | 141 | 202 | 255 | 300 | 301 | 302 | 303 | 304 | 305 | 306 | 307 | 308 | 309 | 310 | 269 | 220 | 163 | 98 | 25 |
| 71 | 140 | 201 | 254 | 299 | 336 | 337 | 338 | 339 | 340 | 341 | 342 | 343 | 344 | 311 | 270 | 221 | 164 | 99 | 26 |
| 70 | 139 | 200 | 253 | 298 | 335 | 364 | 365 | 366 | 367 | 368 | 369 | 370 | 345 | 312 | 271 | 222 | 165 | 100 | 27 |
| 69 | 138 | 199 | 252 | 297 | 334 | 363 | 384 | 385 | 386 | 387 | 388 | 371 | 346 | 313 | 272 | 223 | 166 | 101 | 28 |
| 68 | 137 | 198 | 251 | 296 | 333 | 362 | 383 | 396 | 397 | 398 | 389 | 372 | 347 | 314 | 273 | 224 | 167 | 102 | 29 |
| 67 | 136 | 197 | 250 | 295 | 332 | 361 | 382 | 395 | 400 | 399 | 390 | 373 | 348 | 315 | 274 | 225 | 168 | 103 | 30 |
| 66 | 135 | 196 | 249 | 294 | 331 | 360 | 381 | 394 | 393 | 392 | 391 | 374 | 349 | 316 | 275 | 226 | 169 | 104 | 31 |
| 65 | 134 | 195 | 248 | 293 | 330 | 359 | 380 | 379 | 378 | 377 | 376 | 375 | 350 | 317 | 276 | 227 | 170 | 105 | 32 |
| 64 | 133 | 194 | 247 | 292 | 329 | 358 | 357 | 356 | 355 | 354 | 353 | 352 | 351 | 318 | 277 | 228 | 171 | 106 | 33 |
| 63 | 132 | 193 | 246 | 291 | 328 | 327 | 326 | 325 | 324 | 323 | 322 | 321 | 320 | 319 | 278 | 229 | 172 | 107 | 34 |
| 62 | 131 | 192 | 245 | 290 | 289 | 288 | 287 | 286 | 285 | 284 | 283 | 282 | 281 | 280 | 279 | 230 | 173 | 108 | 35 |
| 61 | 130 | 191 | 244 | 243 | 242 | 241 | 240 | 239 | 238 | 237 | 236 | 235 | 234 | 233 | 232 | 231 | 174 | 109 | 36 |
| 60 | 129 | 190 | 189 | 188 | 187 | 186 | 185 | 184 | 183 | 182 | 181 | 180 | 179 | 178 | 177 | 176 | 175 | 110 | 37 |
| 59 | 128 | 127 | 126 | 125 | 124 | 123 | 122 | 121 | 120 | 119 | 118 | 117 | 116 | 115 | 114 | 113 | 112 | 111 | 38 |
| 58 | 57 | 56 | 55 | 54 | 53 | 52 | 51 | 50 | 49 | 48 | 47 | 46 | 45 | 44 | 43 | 42 | 41 | 40 | 39 |

Figure 7 Example of how a spiral would be "pulled out" of a 20x20 lattice. This spiral accounts for 399 of the 760 possible N pairs, but accounts for all 400 possible S sites.

In order to determine if the source of the bias was from using spirals or from using the (S, N) statistics the spirals were also analyzed using pseudolikelihood. Using pseudolikelihood to analyze the spiral alone (not the entire lattice at once) returned parameter estimates with the same strong bias as using maximum likelihood with the normalizing constant computed from the

(S, N) values. In fact, the parameter estimates for the spiral from both pseudolikelihood and maximum likelihood were very similar and both significantly different from the parameters used to generate the lattices. Thus, the disconnect between analyzing the full lattice vs. a linearized version of the lattice seems to be responsible for the bias in the parameter estimates.

If an explanation for this strong bias could be determined, the bias could be accounted for in the parameter estimates. Until the bias is explained, however, it is unclear whether using spirals to analyze two-dimensional data is an improved or even equivalent method. Further research would explore this question, or would continue to investigate generating the S and N frequencies for two-dimensional lattices instead of only linear lattices.

Works Cited:

1. Gumpertz, Marcia L., Graham, Jonathan M., Ristaino, Jean B. 1997. Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence. Journall of Agricultural, Biological, and Environmental Statistics. 2,2:131-156.

All calculations were performed in R.