# Parameter Estimation in a Hierarchical Model for Species Occupancy

**Rebecca A. Hutchinson**
School of EECS
Oregon State University
Corvallis, OR 97331
rah@eecs.oregonstate.edu

**Thomas G. Dietterich**
School of EECS
Oregon State University
Corvallis, OR 97331
tgd@eecs.oregonstate.edu

## Abstract

In this paper, we describe a model for the relationship between the occupancy pattern of a species on a landscape and imperfect observations of its presence or absence. The structure in the observation process is incorporated generatively, and environmental inputs are incorporated discriminatively. Our experiments on synthetic data compare two methods for training this model under various regularization schemes. Our results suggest that maximizing the expected log-likelihood of the observations and the unknown true occupancy produces parameter estimates that are closer to the truth than maximizing the conditional likelihood of the observations alone.

## 1 Introduction

We consider a problem in which the quantity about which we wish to make inferences is observed imperfectly, with structure in the observation errors. We can encode our knowledge about the observation process by modeling the data generatively, with a latent variable for the quantity of interest that gives rise to observations about its value. In some situations, we also have measurements of variables affecting the latent variable and/or the observation process. We are not interested in modeling the distributions of these variables, so we encode them in the model discriminatively.

This problem occurs in species occupancy modeling, in which the goal is to discover the pattern of occupancy for a species of interest from observations of its presence or absence at randomly sampled locations over a landscape. Since the species might not be detected even when it is present, the sample sites are visited multiple times. In carefully designed studies, the visits are scheduled such that it is reasonable to assume that the species' occupancy does not change over the course of the visits, and that the visits are independent when conditioned on the true occupancy status of the species. The data collected in these studies then contains a detection history for each site, recording the observed presence or absence of the species on each visit. This data is the result of two components: occupancy, which is the biological pattern of interest, and detection, which is a confounding factor. Each component of the model may be accompanied by a set of covariates thought to affect them. For instance, covariates of the occupancy component might include elevation and vegetation type. Covariates of the detection component might include time of day and weather on the day of the visit. The structure of the data is illustrated in Figure 1. Here, $Y_{it}$ is a binary variable recording the presence or absence of the species at site $i$ on visit $t$. $Z_i$ is the true, unobserved, binary occupancy status of the species at site $i$. $X_i$ is a set of covariates for occupancy measured at site $i$, and $W_{it}$ is a set of covariates for detection measured at site $i$ on visit $t$.

In the ecology literature (e.g. [2]), several models of this type are usually fit with different sets of covariates and evaluated according to some model selection criterion. As an alternative, we propose including all covariates in a single model and using regularization to penalize excess complexity.

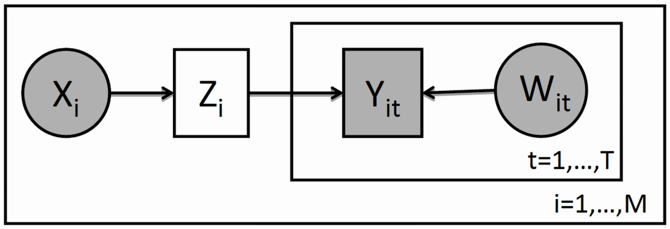

Figure 1: The graphical structure of the species occupancy model. $i$ indexes sites, and $t$ indexes visits. $Y$ is observed data, $Z$ is the latent occupancy, $W$ contains detection covariates, and $X$ contains occupancy covariates.

In this paper, we use synthetic data to investigate two training algorithms for this model, and the regularization trade-offs for each.

## 2 Model Parameterization and Estimation

The model in Figure 1 is commonly parameterized using logistic regressions for the relationship of each component to its covariates, and a zero-inflated binomial for relating occupancy to detection according to the assumptions of the study. More precisely, the distributions are as follows.

$$P(Z_i|X_i,\alpha) = o_i^{Z_i}(1-o_i)^{1-Z_i} \qquad \text{where } o_i = \frac{\exp(\alpha X_i)}{1+\exp(\alpha X_i)} \tag{1}$$

$$P(Y_{it}|Z_i,W_{it},\beta) = (Z_i d_{it})^{Y_{it}}(1-Z_i d_{it})^{1-Y_{it}} \qquad \text{where } d_{it} = \frac{\exp(\beta W_{it})}{1+\exp(\beta W_{it})} \tag{2}$$

We consider two methods for estimating the parameters $\alpha$ and $\beta$: maximizing the conditional log-likelihood $\log(P(Y|X,W,\alpha,\beta))$, or using Expectation Maximization (EM) ([1]) to maximize the expected joint log-likelihood $E[\log(P(Y,Z|X,W,\alpha,\beta))]$.

The conditional log-likelihood is:

$$l(\alpha,\beta|Y,X,W) \quad = \quad \sum_{i=1}^{M}[\log(o_i\prod_{t=1}^{T}[(d_{it})^{Y_{it}}(1-d_{it})^{1-Y_{it}}] + (1-o_i)I(\sum_{t=1}^{T}Y_{it} == 0))] \tag{3}$$

where $I(x)$ returns 1 if and only if $x$ is true. We can use gradient methods to solve for the values of $\alpha$ and $\beta$ that maximize $l$. To regularize the model, we can subtract a penalty function $r(\alpha,\beta)$ from $l$ to form the objective function.

The expected joint log-likelihood is given by the following equation, where the expectation is being taken over the hidden variables given the observed variables, $P(Z|Y,X,W)$:

$$\begin{aligned} Q \quad = \quad & \sum_{i=1}^{M}[P(Z_i=1|Y_{i\cdot},X_i,W_{i\cdot})(\log(o_i)+\sum_{t=1}^{T}(Y_i\log(d_{it})+(1-Y_{it})\log(1-d_{it})))+ \\ & P(Z_i=0|Y_{i\cdot},X_i,W_{i\cdot})(\log(1-o_i)+\sum_{t=1}^{T}Y_{it}(-Inf))] \end{aligned} \tag{4}$$

In the last line, we let the product of 0 and $-Inf$ be 0, so that the last term does not contribute to the likelihood if $Y_{it}$ is 0, and causes the likelihood to be $-Inf$ if $Y_{it}$ is 1. This is consistent with the assumption that if the site is truly unoccupied, the species will never be observed (e.g. misidentified), which is reasonable for expert observers.

To estimate $\alpha$ and $\beta$, the EM algorithm iterates between computing $P(Z|Y,X,W,\alpha,\beta)$ via Bayes rule in the E step and using gradient methods to update $\alpha$ and $\beta$ in the M step. To regularize this model, we use the objective $Q - r(\alpha,\beta)$.

|  | Max Conditional Likelihood |  |  |  |  | EM |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | $\lambda_d$ |  |  |  |  | $\lambda_d$ |  |  |  |  |
| $\lambda_o$ | 0 | 0.0001 | 0.001 | 0.01 | 0.1 | 0 | 0.0001 | 0.001 | 0.01 | 0.1 |
| 0 | 286 | 276 | 283 | 22.2 | 22.1 | 1.62 | 1.62 | 1.61 | 1.63 | 1.25 |
| 0.0001 | 2.50 | 1.93 | 2.89 | 20.7 | 20.7 | 1.71 | 1.61 | 1.70 | 1.63 | 1.25 |
| 0.001 | 3.48 | 1.90 | 1.03 | 8.09 | 11.7 | 1.69 | 1.60 | 1.59 | 1.61 | 1.22 |
| 0.01 | 20.2 | 17.9 | 6.22 | 0.626 | 3.76 | 1.53 | 1.45 | 1.52 | 1.46 | 1.03 |
| 0.1 | 19.3 | 23.1 | 11.5 | 2.43 | 1.25 | 1.03 | 1.03 | 1.02 | 0.995 | 0.473 |

Table 1: Mean squared error, averaged over the 18 parameters, between parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ and the true values that generated the synthetic data, for different regularization and algorithm choices.

## 3 Experimental Results

To compare these two algorithms, along with the regularization trade-offs for each, we ran experiments on synthetic data. The synthetic data consisted of a test set of 250 sites (M=250), each visited twice (T=2). We generated 8 covariates for occupancy, and 8 for detection, only half of which were relevant. That is, 4 of the true $\alpha$ values were drawn from a standard normal distribution, and the other 4 were fixed at 0; $\beta$ was generated in the same way. An intercept term was added to both models. The values of $X$ and $W$ were drawn from standard normal distributions, with ones corresponding to the intercept. The detection histories for the test set were created according to the generative model in Figure 1, and the true values of $Z$ were saved for evaluation purposes, even though $Z$ would not be available in real data. Training sets, also consisting of 2 visits each to 250 sites, were generated similarly, using the same true values for $\alpha$ and $\beta$.

We regularized the two components of the model separately, using the penalty:

$$r(\alpha, \beta) = \lambda_o \frac{1}{2} \sum_{j=2}^{P} \alpha_j^2 + \lambda_d \frac{1}{2} \sum_{k=2}^{Q} \beta_k^2 \tag{5}$$

Note that there are two regularization parameters to tune ($\lambda_o$ and $\lambda_d$) instead of just one, and that we do not regularize the intercept terms. This penalty was used despite the fact that it does not match the true prior distribution on the weights (the ridge penalty corresponds to a Gaussian prior ([4]), and the weights are generated with a zero-inflated Gaussian) because it is easily differentiable.

In the experiments, we let $(\lambda_o, \lambda_d)$ vary over all possible combinations of the values $\{0, 0.0001, 0.001, 0.01, 0.1\}$. For each setting, we generated 30 training sets using the procedure outlined above. Each training set was used to estimate $\hat{\alpha}$ and $\hat{\beta}$ using both algorithms. The resulting estimates were compared with the true parameters, and were also used to predict the true occupancy of the test sites, as well as the observed detection histories. The mean squared error of the parameter estimates and the accuracy on each prediction task was averaged over the 30 training sets to produce a performance measure for each setting of the regularization parameters. The results are presented in Tables 1, 2, and 3. The Bayes optimal classification accuracy for occupancy prediction was 92%, and for detection history prediction was 92.6%.

## 4 Conclusions

Our preliminary results suggest that estimating the model parameters using EM to maximize the joint likelihood of the detection histories and the unobserved variable of interest results in lower mean squared error between the estimates and the truth than maximizing the conditional likelihood of just the detection histories. The parameters recovered by EM also provide a slight improvement on the

|  | Max Conditional Likelihood |  |  |  |  | EM |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | $\lambda_d$ |  |  |  |  | $\lambda_d$ |  |  |  |  |
| $\lambda_o$ | 0 | 0.0001 | 0.001 | 0.01 | 0.1 | 0 | 0.0001 | 0.001 | 0.01 | 0.1 |
| 0 | 90.31 | 90.25 | 90.15 | 90.05 | 89.3 | 90.31 | 90.29 | 90.35 | 90.33 | 90.25 |
| 0.0001 | 90.33 | 90.32 | 90.19 | 90.01 | 89.27 | 90.31 | 90.28 | 90.29 | 90.33 | 90.25 |
| 0.001 | 90.33 | 90.32 | 90.28 | 90.13 | 89.45 | 90.31 | 90.29 | 90.28 | 90.33 | 90.25 |
| 0.01 | 90.24 | 90.35 | 90.37 | 90.20 | 89.81 | 90.32 | 90.28 | 90.32 | 90.33 | 90.27 |
| 0.1 | 89.16 | 89.16 | 89.40 | 88.60 | 88.96 | 90.28 | 90.27 | 90.27 | 90.33 | 90.29 |

Table 2: Accuracy predicting the true occupancy state of the test sites using parameters $\hat{\alpha}$ and $\hat{\beta}$ estimated with different regularization and algorithm choices. The Bayes optimal classification accuracy for this task was 92%.

|  | Max Conditional Likelihood |  |  |  |  | EM |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | $\lambda_d$ |  |  |  |  | $\lambda_d$ |  |  |  |  |
| $\lambda_o$ | 0 | 0.0001 | 0.001 | 0.01 | 0.1 | 0 | 0.0001 | 0.001 | 0.01 | 0.1 |
| 0 | 91.65 | 91.60 | 91.56 | 91.43 | 90.75 | 91.66 | 91.65 | 91.69 | 91.65 | 91.57 |
| 0.0001 | 91.68 | 91.63 | 91.57 | 91.29 | 90.79 | 91.66 | 91.66 | 91.66 | 91.65 | 91.57 |
| 0.001 | 91.66 | 91.63 | 91.55 | 91.52 | 90.84 | 91.68 | 91.67 | 91.65 | 91.65 | 91.57 |
| 0.01 | 91.65 | 91.61 | 91.73 | 91.60 | 90.96 | 91.69 | 91.67 | 91.67 | 91.68 | 91.59 |
| 0.1 | 91.04 | 91.17 | 91.27 | 90.75 | 90.47 | 91.63 | 91.64 | 91.64 | 91.67 | 91.62 |

Table 3: Accuracy predicting detection histories on the test set, using parameters $\hat{\alpha}$ and $\hat{\beta}$ estimated with different regularization and algorithm choices. The Bayes optimal classification accuracy for this task was 92.6%.

classification accuracies we measured (producing higher accuracies in 18/25 cases for occupancy prediction and 23/25 cases for detection history prediction). Note that both training algorithms achieved accuracies very close to the Bayes optimal rate; it is possible that a harder problem would demonstrate a bigger difference between the methods. These results also suggest that EM is less sensitive to the regularization parameters than maximizing the conditional likelihood, at least within the range tested.

The experiments described above are part of a larger body of ongoing work. Firstly, we would like to understand why EM outperforms the conditional likelihood method in our experiments. We are also investigating how these results are affected by various aspects of the data including the true occupancy rate, the true detection rate, the number of visits, the true distribution of the parameters, the number of covariates, and the proportion of relevant covariates. Additionally, we are interested in different regularization penalties, including the elastic net ([5]), in scenarios where the corresponding prior of the penalty may or may not match the true distribution of the weights. In the future, we plan to replace the logistic regressions in the model with tree-based methods ([4]), which have been successful in other species distribution modeling problems ([3]).

## References

[1] A.P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[2] D.I. MacKenzie et al. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, 2006.

[3] J. Elith et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129–151, 2006.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2 edition, 2009.

[5] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320, 2005.