Presence-only data and the EM algorithm

G. Ward,[1,*] **T. Hastie,**[1] **S. Barry,**[2] **J. Elith**[3] **and J.R. Leathwick**[4]

[1] Department of Statistics, Stanford University, CA 94305, U.S.A.

[2] Australian Government Department of Agriculture, Fisheries and Forestry, GPO Box 858, Canberra ACT 2601, Australia

[3] School of Botany, The University of Melbourne, Parkville, Victoria, Australia

[4] National Institute of Water and Atmospheric Research, P O Box 11115, Hamilton, New Zealand

SUMMARY

In ecological modeling of the habitat of a species, it can be prohibitively expensive to determine species absence. Presence-only data consists of a sample of locations with known presences and a separate group of locations sampled from the population, with unknown presences. We propose an EM algorithm to estimate the underlying presence-absence logistic model for presence-only data. This algorithm can be used with any off-the-shelf logistic model. For models with stepwise fitting procedures, such as boosted trees, the fitting process can be accelerated by interleaving expectation steps within the procedure. Preliminary analyses based on sampling from presence-absence records of fish in New Zealand rivers illustrates that this new procedure can reduce both deviance and the shrinkage of marginal effect estimates that occur in the naive model often used in practice. Finally, it is shown that the population prevalence of a species is only identifiable when there is some unrealistic constraint on the structure of the logistic model. In practice, it is strongly recommended that an estimate of population prevalence be provided.

KEY WORDS: Presence-only data; Use-availability data; EM algorithm; Logistic model; Boosted trees

[*] *email:* gward@stanford.edu

1. Introduction

Modeling wildlife habitat selection is important for effective ecological management of a species. Logistic models are typically used to estimate species distribution, ideally based on recorded presences or absences at randomly sampled sites. However, obtaining such presence-absence data is often difficult or expensive, and in practice data arise from alternative sampling mechanisms. Often records of presences are easier and cheaper to obtain than definitive absences. For example, herbaria and museums have extensive historical occurrence records (Elith *et al.*, 2006) and radiotelemetry provides a rich source of range locations for mobile species (Frair *et al.*, 2004). Environmental information can be collected for these recorded presence sites using geographical information system (GIS) technology. Some methods use only these presences to estimate a species range; for example HABITAT (Walker & Cocks, 1991) estimates the range of a species via a convex hull defined in environmental space. However, many methods also require a *background sample* or *pseudo-absences* - a random sample of sites taken from the population of interest. Although the presence or absence of a species at the background sites is unknown, using these data can considerably increase prediction accuracy (Elith *et al.*, 2006). This combination of a sample of locations with known presences, and a background sample of locations from the whole population is what we will refer to as presence-only data. Our definition of the whole or full population refers to all sites in a region of investigation, and should be carefully defined on a case-by-case basis. In our example, studying fish in New Zealand rivers, the population would be all locations along rivers or streams in New Zealand.

The model construction in this paper is based on strict assumptions about the sampling mechanisms. In particular, we assume that the observed presences in the presence-only sample are taken at random from all locations, at a rate proportional to the probability of presence. Additionally, we assume that the background sample is sampled at random from the full population of locations. In practice, this second assumption is often approximately true; GIS provides an easy way to generate environmental covariates for locations chosen at random. However, the presence-only sample is often biased. Records at herbaria or museums are typically ad-hoc records of species occurrence collected by individuals, and may be biased, for example, towards more accessible locations (Reese *et al.*, 2005). Such biases could adversely affect model estimation, and should be investigated on a case-to-case basis.

The modeling of presence-only data in ecology is reviewed in Keating & Cherry (2004), where it is referred to as use-availability data, and Pearce & Boyce (2005). Typically the recorded presences and some pseudo-absences are fit to a logistic model directly; we refer to this as the *naive model*. These pseudo-absences are usually the back-

ground data described above, but e.g. Zaniewski *et al.* (2002) create pseudo-absences by sampling background data that have similar environmental characteristics to sites with recorded presences of other species. If a species is rare, the pseudo-absences will resemble the true absences and the naive model will be close to the true model. However, for want of a better model, the naive model is often used when presences are more common; in this case, we will show that the naive model can be highly biased. An alternate procedure, the exponential model of Manly (2002), justifies the fitting of a naive logistic regression model as a log-linear model of the presence-absence probability. However, this approach does not attempt to estimate the true presence-absence logistic model, and has other recorded failings (Keating & Cherry, 2004).

Lancaster & Imbens (1996) provide an exact method for estimating the a logistic regression model with presence-only data. It uses direct maximization of the presence-only likelihood to estimate the linear predictor for the presence-absence logistic regression model. However, this technique is rarely used in practice as there is no easily-available implementation and convergence problems have been reported. Our method indirectly maximizes the presence-only likelihood in a way that is more robust, can easily incorporate more flexible models and is straightforward to implement using existing statistical software.

Maxent or maximum entropy modeling (Phillips *et al.*, 2006) approaches the problem from an unconditional perspective. Maxent directly estimates a species distribution, whereas logistic regression models the conditional probabilities of presence at a location. The entropy of the distribution is maximized with respect to constraints on the similarity of model and empirical statistics. Additionally, a recent extension (Dudík *et al.* 2005) has tried to model and remove the effect of any bias in the sampling distribution of the presences. Although initial experiments on real data indicate mixed results, the explicit modeling and removal of sampling bias is of great interest for the presence-only problem. Both Maxent and naive conditional models have been shown to perform well in practice (Elith *et al.*, 2006). However, in this paper we concentrate on improving the conditional model, using modern yet well-established and highly flexible procedures, and directly addressing the presence-only problem.

Within ecology the presence-only problem arises in paradigms with subtle distinctions. In "use-availability" modeling, we are interested in the relative frequency of use of an area by wide-ranging animals; for less mobile species, "presence-absence" is modeled. However, for both these cases it is possible to arrive at a unique definition. We define the probability of presence of a species as measured across a certain time window. For non-mobile species, this time window is irrelevant, up to the lifetime of the species. Typically, for mobile species, the window is implicitly defined by the sampling

mechanism, for example by the length of time a site was observed, or by the discretization of radiotelemetry data. Any reference to a probability of presence in this paper thus implicitly refers to such a time window. In addition, to avoid confusion, in this paper we refer to the presence-only and background samples as *naive presences* and *naive absences* respectively, in reference to the naive procedure of fitting a logistic regression model directly to these data.

In Sections 2.1 to 2.4 we motivate and develop the application of the EM algorithm to the presence-only problem when the population prevalence is known. An adjustment to the procedure is proposed for stepwise modeling procedures (Section 2.5) and the inadequacies of the naive model is illustrated in Section 2.6. The method is illustrated for a species of New Zealand eel in Section 3. Finally, in Section 4 we show that the simultaneous estimation of the population prevalence along with the model is only possible given certain modeling assumptions, which we believe are too stringent to justify.

2. Models and Procedures

A common aim of ecological studies is to model the probability that a species of interest is present, $y = 1$, at a location with covariates $\mathbf{x}$, $\mathbb{P}(y = 1|\mathbf{x})$. This probability is usually modeled via its logit

$$\text{logit}\, \mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}) \;\; \Rightarrow \;\; \mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1 + e^{\eta(\mathbf{x})}} \tag{1}$$

where $\eta(\mathbf{x})$ can be linear in $\mathbf{x}$ (as in logistic regression) or a non-linear function e.g. GAMs (Hastie & Tibshirani, 1990) or boosted trees (Friedman, 2001). In studies where the true presences are known, and for a simple random sample of sites, these models are fit using established methods.

However, for the presence-only problem we do not know the true presences y. Instead we know an observed or naive presence, $z = 1$, and the background data or naive absence $z = 0$. There are two reasons why we cannot use the established methods directly on the naive presences/absences z. Firstly, there is missing information about the true presence or absence at each location in the background sample. Secondly, even if the true presences/absences are known, we do not have a simple random sample of sites; the extra sample of presences mean these data have a higher proportion of presences than in the population. In Section 2.1 we investigate how to overcome the second problem using a modified case-control methodology, which in Section 2.2 we use to find the easily-optimized *full likelihood* for the true and naive presences. We then show that the *observed likelihood*, for the naive presences only, is not straightforward to maximize directly (Section 2.3), but that an application of the EM algorithm to the full likelihood helps us overcome the problem of missing information in a simple and elegant way

(Section 2.4). In Section 2.5 we see how computational efficiency of the EM algorithm can be increased for stepwise logistic modeling procedures. Finally, we see that a naive fitting of a logistic model gives a biased estimate of the presence-absence model, while the EM procedure gives approximately unbiased estimates (Section 2.6). In this section we assume that we know the overall population prevalence of the species π. Later we look at an example of sensitivity analysis of uncertainty in π, and illustrate under what conditions we can also estimate π directly from these data.

2.1 *Modified Case-Control Sampling*

In this section we assume that the true presences and absences $\mathbf{y}$ are known for all the data. Using these data we will illustrate how to estimate the model of interest (1) via a modified case-control methodology, which will then be used in the EM algorithm described later in the paper.

Traditional case-control sampling, e.g. McCullagh and Nelder (1989, p111), is a retrospective sampling technique whereby cases and controls (or equivalently presences and absences) are sampled separately, at different and unknown rates. We can write these rates as $\gamma_1 = P(s = 1|y = 1)$ and $\gamma_0 = P(s = 1|y = 0)$ where $s = 1$ indicates that an observation is in the sample. We assume that γ_1 and γ_0 are both independent of the covariates $\mathbf{x}$. If there are n_1 cases and n_0 controls in the sample, and if N is the size of the population, then

$$\gamma_1 = \frac{n_1}{\pi N} \quad \text{and} \quad \gamma_0 = \frac{n_0}{(1-\pi)N},$$

where π is the proportion of cases or presences in the population. To address the different sampling rates, we model the case-control probability $P(y = 1|s = 1, \mathbf{x})$ conditional on the event that an observation is in the sample.

Using Bayes rule we have

$$\begin{aligned}
\mathbb{P}(y = 1|s = 1, \mathbf{x}) &= \frac{\mathbb{P}(s = 1|y = 1, \mathbf{x})\, \mathbb{P}(y = 1|\mathbf{x})}{\mathbb{P}(s = 1|y = 0, \mathbf{x})\, \mathbb{P}(y = 0|\mathbf{x}) + \mathbb{P}(s = 1|y = 1, \mathbf{x})\, \mathbb{P}(y = 1|\mathbf{x})} \\
&= \frac{\gamma_1 e^{\eta(\mathbf{x})}}{\gamma_0 + \gamma_1 e^{\eta(\mathbf{x})}} \\
&= \frac{e^{\eta^*(\mathbf{x})}}{1 + e^{\eta^*(\mathbf{x})}}
\end{aligned} \tag{2}$$

where $\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log(\gamma_1/\gamma_0)$. Plugging in the sampling rates gives

$$\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log\left(\frac{n_1}{n_0}\right) - \log\left(\frac{\pi}{1-\pi}\right). \tag{3}$$

Thus the case-control model is the true model (1), but where the true proportions of π and $1 - \pi$ are replaced by the empirical priors of n_1 and n_0. Even when the population prevalence π is unknown, η is identifiable up to a constant.

In our example, the true presences are split between the naive presences and absences, so the sampling rates calculation is slightly different. Let n_p be the number of naive presences and n_u be the number of observations in the background data. Then

$$\mathbb{P}(y = 1|s = 1) = \frac{n_p + \pi n_u}{n_p + n_u} \quad \text{and} \quad \mathbb{P}(y = 1|s = 0) = \frac{(1-\pi)n_u}{n_p + n_u}$$

are the expected proportion of presences and absences respectively in the whole data set. Using Bayes rule,

$$\gamma_1 = \mathbb{P}(s = 1|y = 1) = \frac{\mathbb{P}(y = 1|s = 1)\,\mathbb{P}(s = 1)}{\mathbb{P}(y = 1)} = \frac{n_p + \pi n_u}{\pi\,(n_p + n_u)}\,\mathbb{P}(s = 1)\,, \text{ and}$$

$$\gamma_0 = \mathbb{P}(s = 1|y = 0) = \frac{\mathbb{P}(y = 0|s = 1)\,\mathbb{P}(s = 1)}{\mathbb{P}(y = 0)} = \frac{(1-\pi)n_u}{(1-\pi)(n_p + n_u)}\,\mathbb{P}(s = 1).$$

and the offset in the modified case-control model is

$$\log\left(\frac{\gamma_1}{\gamma_0}\right) = \log\left(\frac{n_p + \pi n_u}{(1-\pi)n_u}\right) - \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{n_p + \pi n_u}{\pi n_u}\right).$$

This is equivalent to (3) where $n_1 = n_p + \pi n_u$ is now the expected number of true presences given the naive sample sizes. As we are assuming that π is known, the true model η can be estimated outright.

To summarize, if we knew the true presences/absences for all observations, we could estimate the model of interest (1) by fitting the usual model to these data and then subtracting the constant $\log\left(\frac{n_p+\pi n_u}{\pi n_u}\right)$ from $\eta(\mathbf{x})$. This will be the maximization step in an EM algorithm, but to show this we first need to consider the likelihoods of the full and observed data.

2.2 *The Full Likelihood*

Continuing the pretense that we know all presences/absences, the next step is to write out the likelihood. We note that as well as the true presences $\mathbf{y}$, we know the naive presences $\mathbf{z}$. Thus the likelihood is of the joint probabilities of $\mathbf{y}$ and $\mathbf{z}$ conditional on X:

$$L(\eta|\mathbf{y}, \mathbf{z}, X) = \prod_i \mathbb{P}(y_i, z_i|s_i = 1, \mathbf{x}_i) = \prod_i \mathbb{P}(y_i|s_i = 1, \mathbf{x}_i)\,\mathbb{P}(z_i|y_i, s_i = 1, \mathbf{x}_i) \tag{4}$$

This factors into the case-control probabilities $\mathbb{P}(y_i|s_i = 1, \mathbf{x}_i)$ and the sampling probabilities for the naive presences and absences, conditional on the true presence, $\mathbb{P}(z_i|y_i, s_i = 1, \mathbf{x}_i)$. Note that

$$\mathbb{P}(z_i = 0|y_i = 0, s_i = 1, \mathbf{x}_i) = 1$$

because there are no true absences in the sample of naive presences. Additionally, given the true presence or absence, the sampling of the naive presences was independent of $\mathbf{x}$

so

$$\mathbb{P}(z_i = 1|y_i = 1, s_i = 1, \mathbf{x}_i) = \frac{\mathbb{P}(z_i = 1, y_i = 1|s_i = 1)}{\mathbb{P}(y_i = 1|s_i = 1)} = \frac{n_p}{n_p + \pi n_u}$$

and $\mathbb{P}(z_i = 0|y_i = 1, s_i = 1, \mathbf{x}_i) = \frac{\pi n_u}{n_p + \pi n_u}$. None of these probabilities depend on η, so they only appear in the likelihood as a constant. However, we will re-visit these in Section 4 when we consider simultaneous estimation of π.

Thus, when we know π, the likelihood depends only on the case-control probabilities:

$$L(\eta|\mathbf{y}, \mathbf{z}, X) \propto \prod_i \mathbb{P}(y_i|s_i = 1, \mathbf{x}_i) \tag{5}$$

This is the usual case-control likelihood. To estimate η, we fit a model to the true presences $\mathbf{y}$ to obtain $\hat{\eta}^*$, and then subtract the constant $\log\left(\frac{n_p + \pi n_u}{\pi n_u}\right)$. Although we do not know the true presences for the background data, in the next few sections we will see that we can still estimate η this way, using an application of the EM algorithm.

2.3 *The Likelihood of the Observed Data*

In this section we return to looking at the observed data: the naive presences and absences. To estimate η for the presence-absence model (1) we wish to maximize the likelihood for the presence-only data, with respect to η. In this section we generate the likelihood for the presence-only data to show that it is not easy to optimize for general η. This motivates the use of the EM algorithm on the full likelihood.

Let $z = 1$ indicate a naive presence and $z = 0$ indicate a naive absence, i.e. the background sample. The probability of the presence-only data being a naive presence is $\mathbb{P}(z = 1|s = 1, \mathbf{x})$, where $s = 1$ indicates that the location has been sampled. This can be written in terms of the case-control probabilities and sampling probabilities of the z given y:

$$\begin{aligned}\mathbb{P}(z = 1|s = 1, \mathbf{x}) =& \mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x})\mathbb{P}(y = 1|s = 1, \mathbf{x}) \\ &+ \mathbb{P}(z = 1|y = 0, s = 1, \mathbf{x})\mathbb{P}(y = 0|s = 1, \mathbf{x}).\end{aligned}$$

We are assuming that the naive presences and absences are simple random samples from the population of presences and the total population respectively. Given this, the sampling probabilities $\mathbb{P}(z|y, s = 1, \mathbf{x})$ are independent of $\mathbf{x}$. Further, $\mathbb{P}(z = 1|y = 0, s = 1) = 0$ because all true absences in the data must be in the background sample. Using

calculations from Section 2.1,

$$\begin{aligned}\mathbb{P}(z=1|y=1,s=1) &= \frac{\mathbb{P}(z=1,y=1|s=1)}{\mathbb{P}(y=1|s=1)} \\ &= \frac{n_p/(n_p+n_u)}{(n_p+\pi n_u)/(n_p+n_u)} \\ &= \frac{n_p}{n_p+\pi n_u}.\end{aligned}$$

Remembering that $\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log\left(\frac{n_p+\pi n_u}{\pi n_u}\right)$, from (2.3) we have

$$\begin{aligned}\mathbb{P}(z=1|s=1,\mathbf{x}) &= \frac{n_p}{n_p+\pi n_u}\,\frac{e^{\eta^*(\mathbf{x})}}{1+e^{\eta^*(\mathbf{x})}} + 0 \\ &= \frac{\frac{n_p}{\pi n_u}e^{\eta(\mathbf{x})}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x})}}.\end{aligned} \tag{6}$$

The probability of a naive presence is a multiple of the true probability of a presence (recorded or not), with an adjusted intercept of η; it is the combination of these two adjustments that make the presence-only problem difficult to solve.

The observed likelihood for the presence-only problem is the product of the probabilities of the naive presences $\mathbf{z}$:

$$\begin{aligned}L(\eta|\mathbf{z},X) &= \prod_i \mathbb{P}(z_i|s_i=1,\mathbf{x}_i) \\ &= \prod_i \left(\frac{\frac{n_p}{\pi n_u}e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right)^{z_i}\left(\frac{1+e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right)^{1-z_i}.\end{aligned} \tag{7}$$

This likelihood can be maximized directly in certain situations, e.g. when η is linear (Lancaster & Imbens, 1996), but for more general η, direct maximization is infeasible. In the next section we will show how the EM algorithm provides an elegant solution.

2.4 *The EM Algorithm*

The EM algorithm (Dempster *et al.*, 1977) is a technique for maximizing a likelihood where there are some latent or unknown variables. The likelihood is maximized by alternating between expectation and maximization steps. The expectation of the log-likelihood is taken with respect to the parameter estimates obtained in the previous step. Then that expected log likelihood is maximized to re-estimate the parameters. For the presence-only problem, we know the naive presences $\mathbf{z}$ and the latent variables are the true presences $\mathbf{y}$. The rest of this section develops the EM algorithm for this case; the procedure is summarized in Figure 1.

1. Chose initial estimates $\hat{y}_i^{(0)} = \pi$ for $i \in \mathbf{U}$.

2. Repeat until convergence:

   - *Maximization step:*
     - Calculate $\hat{\eta}^{*(k)}$ by fitting either

       (1) a logistic model of $\hat{\mathbf{y}}_{\mathbf{U}}^{(k-1)}$ on X

       (2) a logistic model of $\begin{pmatrix} \mathbf{1}_P \\ \mathbf{1}_{\mathbf{U}} \\ \mathbf{0}_{\mathbf{U}} \end{pmatrix}$ on $\begin{pmatrix} X_P \\ X_{\mathbf{U}} \\ X_{\mathbf{U}} \end{pmatrix}$ with weights $\begin{pmatrix} \mathbf{1}_P \\ \hat{\mathbf{y}}_{\mathbf{U}}^{(k-1)} \\ \mathbf{1}_{\mathbf{U}} - \hat{\mathbf{y}}_{\mathbf{U}}^{(k-1)} \end{pmatrix}$

     - Calculate $\hat{\eta}^{(k)} = \hat{\eta}^{*(k)} - \log(\frac{n_p + \pi n_u}{(1-\pi)n_u}) - \log(\frac{\pi}{1-\pi})$.
   - *Expectation step:* For $i \in \mathbf{U}$,

$$\hat{y}_i^{(k)} = \mathbb{E}[y_i | \hat{\eta}^{(k)}] = \frac{e^{\hat{\eta}^{(k)}} + 1}{1 + e^{\hat{\eta}^{(k)}} + 1}.$$

Figure 1. *Estimation of the contaminated case-control model via the EM algorithm, when π is known.*

The full likelihood of both $\mathbf{z}$ and $\mathbf{y}$ (5) is

$$\begin{aligned} L(\eta | \mathbf{y}, \mathbf{z}, X) &\propto \prod_i \mathbb{P}(y_i | s_i = 1, \mathbf{x}_i) \\ &= \prod_i \mathbb{P}(y_i = 1 | s_i = 1, \mathbf{x}_i)^{y_i} \mathbb{P}(y_i = 0 | s_i = 1, \mathbf{x}_i)^{1-y_i} \end{aligned} \tag{8}$$

and so the log likelihood is

$$\log L(\eta | \mathbf{y}, \mathbf{z}, X) = \sum_i \big(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \big) + \text{ constant}$$

where $p_i = \mathbb{P}(y_i = 1 | s_i = 1, \mathbf{x}_i)$. As the log likelihood is linear in the y_i, the expectation step at the kth iteration of the EM algorithm is to replace each y_i with its expectation conditional on the estimated model from the previous step $\eta^{(k-1)}$. Denoting this expectation $\hat{y}_i^{(k)}$ we have

$$\hat{y}_i^{(k)} = \mathbb{E}[y_i | \eta^{(k-1)}, \mathbf{x}_i] = \mathbb{P}(y_i = 1 | \eta^{(k-1)}, \mathbf{x}_i) = \frac{e^{\eta^{(k-1)}(\mathbf{x}_i)}}{1 + e^{\eta^{(k-1)}(\mathbf{x}_i)}} \text{ for } i \in \mathbf{U},$$

where $\mathbf{U}$ is the set of unlabeled data, and $\hat{y}_i^{(k)} = 1$ for $i \in P$, the set of naive presences.

The maximization step involves estimation of a case-control logistic model as described in Section 2.1. Any logistic model fitting procedure can be used, as long as that procedure can handle either weights, or non-integer responses. That is because the $\hat{y}_i^{(k)}$,

for $i \in \mathbf{U}$, lie in the interval $[0, 1]$ and are not binary. If the procedure can handle non-integer responses, then the maximization step is straightforward. Fit $\hat{\mathbf{y}}^{(k)}$ to X, to obtain $\eta^{*(k)}$. Then $\eta^{(k)} = \eta^{*(k)} - \log(\frac{n_p + \pi n_u}{\pi n_u})$.

If the fitting procedure can handle weights but not non-integer responses, the maximization step involves fitting an augmented $\hat{\mathbf{y}}^{(k)}$ to an augmented X. To see this, define weights $w_{k,i} = \hat{y}_i^{(k)}$ and $w^*_{k,i} = 1 - \hat{y}_i^{(k)}$ so that

$$\mathbb{E}[\log L(\eta; \mathbf{y}, \mathbf{z}, X)|\eta^{(k)}] = \sum_{i \in P} \log p_i + \sum_{i \in \mathbf{U}} w_{k,i} \log p_i + \sum_{i \in \mathbf{U}} w^*_{k,i} \log (1 - p_i)$$

and it is apparent that this is equivalent to a weighted logistic model with $n_p + 2n_u$ observations, of which $n_p + n_u$ observations have an outcome of 1, and n_u observations have an outcome of 0. The locations $i \in \mathbf{U}$ now appear twice, with outcomes 1 and 0 and corresponding weights $\hat{y}_i^{(k)}$ and $1 - \hat{y}_i^{(k)}$. More concretely, the maximization step is fitting:

$$\text{a logistic model of } \begin{pmatrix} \mathbf{1}_P \\ \mathbf{1}_\mathbf{U} \\ \mathbf{0}_\mathbf{U} \end{pmatrix} \text{ on } \begin{pmatrix} X_P \\ X_\mathbf{U} \\ X_\mathbf{U} \end{pmatrix} \text{ with weights } \begin{pmatrix} \mathbf{1}_P \\ \hat{\mathbf{y}}_\mathbf{U}^{(k-1)} \\ \mathbf{1}_\mathbf{U} - \hat{\mathbf{y}}_\mathbf{U}^{(k-1)} \end{pmatrix}$$

where the rows of X_P are the rows i of X such that $i \in P$, the elements of $\hat{\mathbf{y}}_\mathbf{U}^{(k-1)}$ are the elements i of $\hat{\mathbf{y}}^{(k-1)}$ such that $i \in \mathbf{U}$, and $\mathbf{1}_\mathbf{U}$ and $\mathbf{0}_\mathbf{U}$ are vectors of ones and zeros of length $|\mathbf{U}|$. After fitting the model, the resulting $\eta^{*(k)}$ must be adjusted as before to obtain $\eta^{(k)}$.

We have seen that the EM algorithm works by alternately fitting the case-control logistic model, adjusting the intercept and then calculating the expectations of the latent y_i for $i \in \mathbf{U}$. Theoretical results in Dempster *et al.* (1977) guarantee that the maximum likelihood estimate for the extended case-control setup is attained. All that is required are suitable starting values for the latent $\mathbf{y}$; using $\hat{y}_i^{(0)} = \pi, \ \forall i \in \mathbf{U}$, so that the sample mean for the background data equals the population mean, ensures that the case-control offset adjustment is correct. Results of the EM procedure applied to a simple logistic regression example are illustrated in Figure 3; the estimates are approximately unbiased even as π increases.

2.5 *Stepwise procedures and the EM method*

Recent work in ecological modeling has seen a move toward more flexible models, such as generalized additive models (GAM) and boosted trees (e.g. Leathwick *et al.*, 2005 and Leathwick *et al.*, in review). Both GAM and boosted trees can be the logistic model of choice in the maximization step of this EM algorithm (Figure 1). However the stepwise nature of the boosted trees model suggests an improvement; interleaving expectation steps with every few boosting steps will decrease the time to convergence of the EM

algorithm. In practice this decrease can be an order of magnitude. Additionally, this interleaving is straightforward to implement using the existing boosted trees package GBM in R (Ridgeway, 2004). Although we illustrate the procedure for a boosted trees model, this could be implemented for any stepwise procedure.

The algorithm follows the same steps as in Figure 1, except that when the M-step requires fitting a logistic model, instead of fitting a whole boosted trees model, we take the existing boosted trees model, update the weights in the model, and then continue to run the fitting procedure to add some more trees. As adding single trees is relatively inexpensive, we run the model to add say 10 or 50 extra trees at each maximization step - this parameter may be adjusted to optimize the fit.

The optimal number of trees is chosen in a similar way to a regular boosted trees model; by minimizing a test-set or cross-validation error. However, we have to be careful here as any built-in error calculations would be based on the estimates $\hat{y}^{(k)}$ from the previous iteration of the fitted model. Instead we use the observed deviance, which is -2 times the log of the observed likelihood given in (7). The observed deviance is calculated using predictions based on the case-control adjusted $\eta^{(k)}$, and the naive presences $\mathbf{z}$ for a separate test set or for cross-validation sets. Once the optimal model size has been chosen, predictions and marginal plots of the effects of each variable or groups of variables can be calculated as usual, via the built-in functions of the boosted trees package.

2.6 *The Naive Logistic Model*

In recent literature (e.g. Ferrier *et al.*, 2002) one approach to the presence-only problem has been to fit a logistic regression directly to the naive presences and absences, ignoring the "contamination" of the absences. We will refer to this as a naive logistic regression and show that it can result in severely biased estimates of η. Let η_{naive} be the linear predictor of the naive logistic model on the presence-only data. We will start by calculating a case-control adjustment for the naive model, and then show how the adjusted η_{naive} is related to the true η.

Fitting a logistic model directly to the naive presences is doubly problematic. As well as ignoring the "contamination" of the absences, the presences and absences are not sampled proportionally to their prevalence in the population. To obtain the best estimate of η from the naive model, we can do a case-control adjustment to η_{naive} (e.g. Mace *et al.*, 1996). To do this, we assume that the data are random samples of size n_p and n_u from the populations of presences and absences respectively. Then from (3), the case-control offset for the naive model is

$$\log\left(\frac{\gamma_1}{\gamma_0}\right) = \log\left(\frac{n_p}{n_u}\right) - \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{(1-\pi)n_p}{\pi n_u}\right).$$

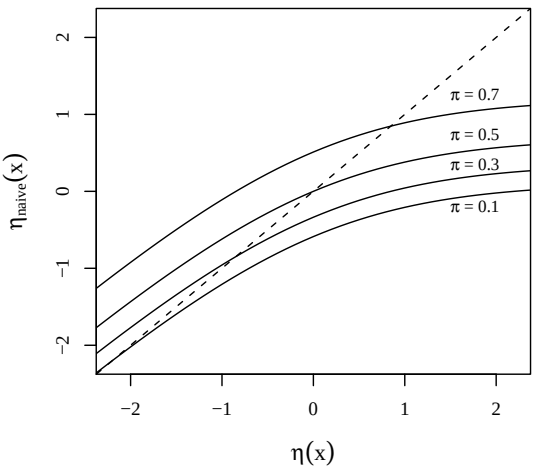


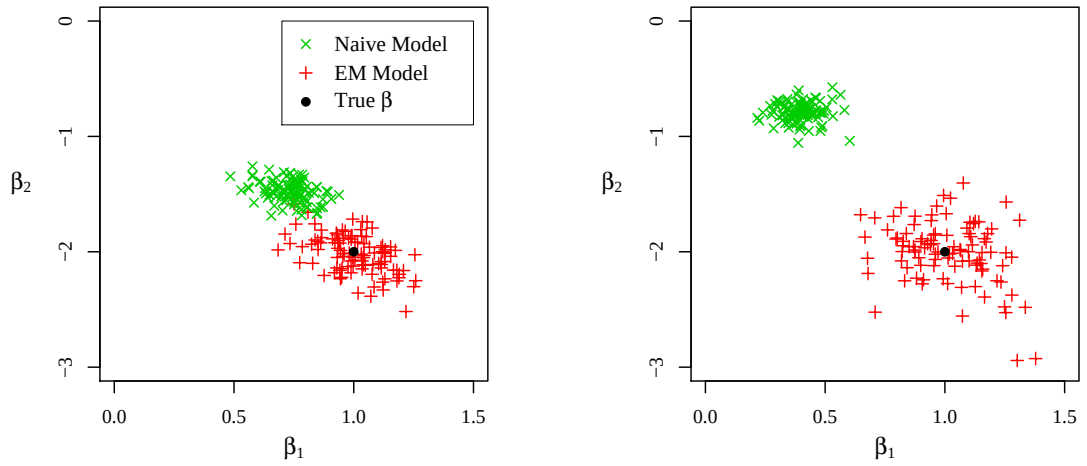
Figure 2. *The case-control adjusted η_{naive} from the naive logistic regression is an increasing but non-linear function of the linear predictor from the true model of interest, η, and the population prevalence π.*

We can write out this case-control adjusted η_{naive} in terms of the true η of interest:

$$\begin{aligned}
\eta_{\text{naive}}(\mathbf{x}) - \log\left(\frac{(1-\pi)n_p}{\pi n_u}\right) &= \text{logit}\left(\mathbb{P}(z=1|s=1,\mathbf{x})\right) - \log\left(\frac{(1-\pi)n_p}{\pi n_u}\right) \\
&= \log\left(\frac{\frac{n_p}{\pi n_u}e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}}\right) - \log\left(\frac{(1-\pi)n_p}{\pi n_u}\right) \\
&= -\log(1-\pi) - \log\left(e^{-\eta(\mathbf{x})}+1\right)
\end{aligned}$$

The naive linear predictor η_{naive} is increasing in the true η, but nonlinearly (Figure 2). Although the true and naive predictors are similar up to a constant for $\eta \ll 0$, when $\eta > 0$ the naive model considerably underestimates the rate at which η is increasing. In particular, the naive linear predictor is bounded above: $\eta_{\text{naive}} \leq -\log(1-\pi)$. Hence the estimated probabilities for the naive model will be underestimated for locations with higher probabilities of presence.

In practice, this bias has considerable effect on logistic regression estimates: if the true η is linear in $\mathbf{x}$, i.e. $\eta(\mathbf{x}) = \mathbf{x}^T\beta$, then any naive logistic regression model must be biased. In particular, the estimates for the naive model β_{naive} will tend to underestimate the slopes β. This is illustrated in simulations of presence-only data (Figure 3). Although the estimates for the EM model are approximately unbiased, the naive model estimates are shrunk to zero, with more shrinkage for larger π. However, note that even for $\pi = 0.1$, the naive estimates are noticeably different from the truth. It is easy to see how this shrinkage may impact variable selection, e.g. important variables may not appear statistically significant.



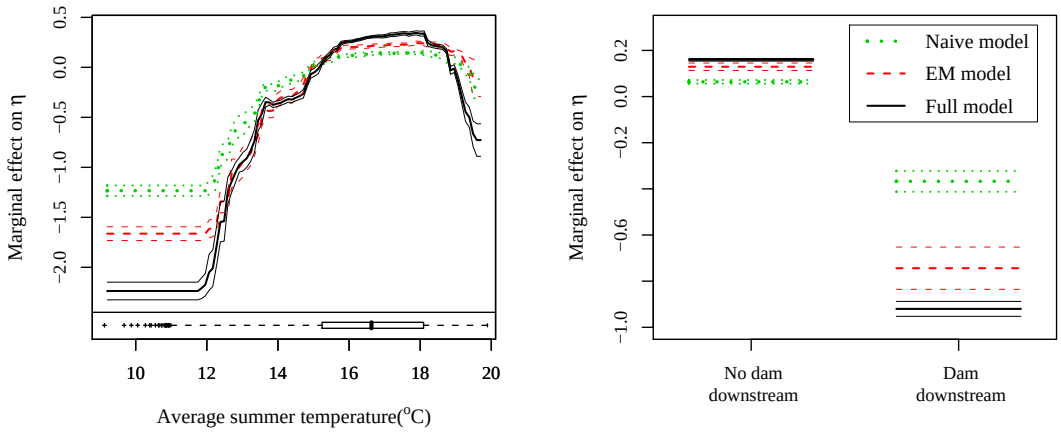
(a) π=0.1 (b) π=0.4

Figure 3. *Parameter estimates for the naive logistic regression model are biased toward the origin, with increased bias for larger π. The EM procedure produces approximately unbiased estimates, but with increasing variance for larger π. These estimates are from 100 simulations of presence-only data, generated from the model $\eta(\mathbf{x}) = \alpha + x_1\beta_1 + x_2\beta_2$, where $\beta_1 = 1$ and $\beta_2 = -2$. The $\mathbf{x}$ are i.i.d. standard normals and $n_p = 300$ and $n_u = 1000$.*

3. Example: The Longfin Eel

Assessing models based on presence-only data is difficult, because there is typically no validation data with known presences and absences. Although simulated data are a useful tool, they typically do not reflect the noise and complex structure inherent in ecological data. To overcome these issues, we have generated presence-only data sets from presence-absence data of diadromous fish in the rivers of New Zealand (more details are given in Leathwick *et al.*, 2005 and in the acknowledgements). In particular we looked at the Longfin Eel *Anguilla dieffenbachii* which has a high prevalence, occurring at 51.3% of all locations sampled. To reduce spurious effects, we repeated the presence-only data generation 10 times, with different random seeds; the results provided are amalgamated across these repetitions. There are 21 environmental covariates describing conditions at the sampled site as well as up- and downstream. This includes some variables omitted from Leathwick *et al.* (2005), in particular the presence of a dam downstream, as this provides a clear illustration of the difference in the naive and EM models. In the full data set, eels were present in 20% of sites with a dam downstream.

The presence-only samples were generated according to the sampling assumptions set out in this paper. Ideally the naive absences should be a random sample from all river locations in New Zealand; here we assume that the fish data set consists of such a sample, so we can compare performances of different models. Hence the sample of



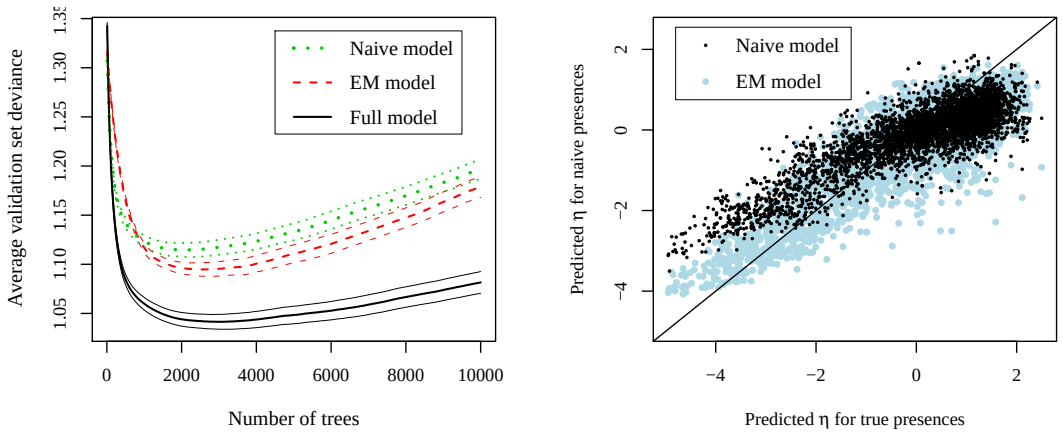
(a) Average summer air temperature (^{o}C)

(b) Presence of a dam downstream

Figure 4. *The EM model provides a better approximation than the naive model of the estimated effects of summer temperature and a downstream dam on the presence of the Longfin Eel. The plots indicate the marginal effect of each variable on η for the EM model, the naive model and for the full model based on the true presences. The three lines for each of the models are the mean across the 10 sampling repetitions, along with ± 1 standard error. The boxplot at the bottom of the left plot indicates the distribution of the summer temperature across all locations in the sample; 15% of these sites had a dam downstream.*

naive absences was generated by sampling randomly from all observations. Then the naive presences sample was generated by sampling from the remaining presence locations. Thus no location could occur in both samples, although in practice this is not a requirement. A random sample of one quarter of the data were set aside as a validation set and the rest were used to train the models. These training sets contained around 4,400 and 1,300 naive absences and presences respectively. The validation set contained one third of these numbers, of which around 1,200 were true presences and 700 were true absences.

We fitted a boosted trees model to these data, using an EM algorithm (Figure 1), which we will call the *EM model*, with $\pi = 0.513$. This is compared to the *naive model*, fitting boosted trees directly to the presence-only data with a case-control adjustment to enable a fair comparison with the EM model. Additionally, as a measure of an optimal result, we calculated the *full model*, a boosted trees model calculated on the same data, but using the full knowledge of the true presences and absences. A total of 10,000 trees were fitted for each model, each with a depth of 4, allowing four-way interaction terms. In practice the optimal number of trees for a model would be chosen by minimizing a prediction set deviance using the observed likelihood (7). However, here the deviance was calculated using the full likelihood (5) of the true presences, so that the EM and naive models could be compared with the full model. As the validation set was not a



(a) Validation set deviance

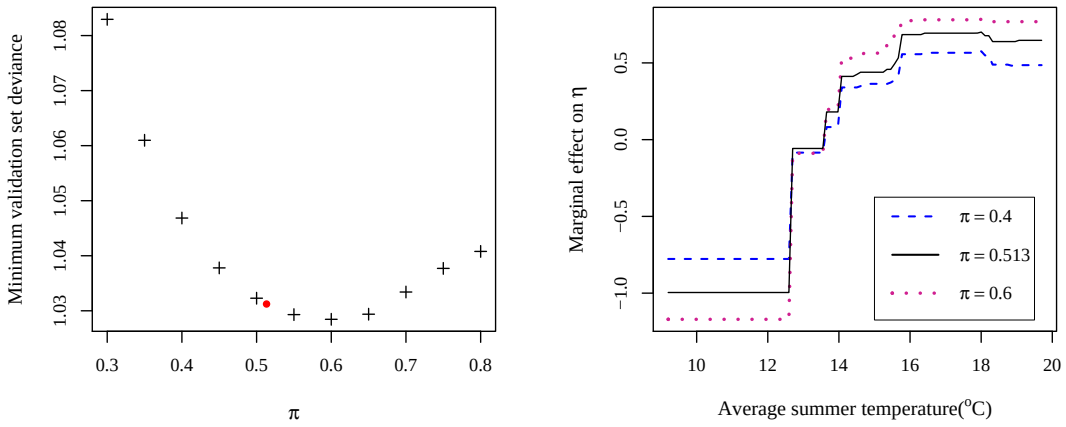
(b) $\hat{\eta}$ for the naive and true presences

Figure 5. *The EM model has a lower validation set deviance and less shrinkage in η than the naive model, when predicting the presence of the Longfin Eel. The left plot indicates the average deviance (calculated from the full likelihood of the true presences) for the validation data for the EM model, the naive model and for the full model based on the true presences, by number of trees in the model. The average deviance is across data in each sampling repetition; the three lines indicate the mean and ± 1 standard error of these averages across the 10 repetitions. The right plot compares the predicted η for the best fitting EM and naive models, based on the presence-only data, versus the best fitting model based on the true presences. These data are from one of the 10 repetitions, although the pattern is replicated across repetitions.*

random sample from the true population, a further case-control adjustment was made in calculating deviances. In general, the full and observed likelihoods generate similar shaped curves that are minimized at a similar number of trees. The naive model required the fewest trees (Figure 5), while the EM models were optimized at around 3,000 trees, similar to the optimal number of trees for the model based on the true presences.

The two most important predictors in modeling the true presences were the summer temperature and the presence of a dam downstream. Figure 4 illustrates that the naive model tends to underestimate the range of the effect that each of these predictors has on η. This is particularly noticeable in the binary downstream dam variable, and echoes the pattern in Figure 3 that the naive logistic regression model shrinks the parameter estimates toward zero. This leads to the effect seen in Figure 5, that the predicted η for the naive model tend to be shrunk to zero in comparison to the model based on the true presences.

The performance of the EM model lies somewhere between that of the naive and true models. There is some shrinkage of the marginal effects and estimated ηs, though it is not as pronounced as for the naive model. This occurs because the presence probabilities imputed at the E-steps are estimates of the true model, thus introducing an



(a) Minimum validation set deviance

(b) Marginal effect on η

Figure 6. *A sensitivity analysis for π indicates that the minimum validation set deviance for the EM model is smallest for $\pi \approx 0.6$. The effect on η of average summer temperature has a consistent shape across all π, but the estimated effect magnitude increases with π. The validation set deviance is calculated using the presence-only likelihood, so is not comparable with Figure 5, and the minimum validation deviance is given to indicate the best fit for that value of π.*

averaging effect on the estimated presences/absences for the background data. Figure 5 illustrates that the EM model also has lower deviance than the naive model (calculated on a validation set where the true presences and absences are known). In this case, with $\pi = 0.513$, the EM model considerably reduces the excess in deviance over the model based on the true presences, compared to the naive model. Although these results are exploratory, these effects were seen across other species of fish with $\pi > 0.2$, but with less of an effect for smaller π. It should be noted, however, that these results were for data that were sampled according to the assumptions set out in this paper; in practice, sampling of presences may be very ad-hoc.

Finally, we ran a sensitivity analysis of the model, for one of the 10 sampling repetitions, to determine the effect of differing prior beliefs of π. The EM model was fitted using a range of different πs between 0.3 and 0.8, in increments of 0.05. The validation set deviance was calculated for each model, using the observed likelihood (7), and the minimum deviance recorded. Figure 6(a) illustrates that these minimum deviances are themselves minimized around $\pi = 0.6$, with low minimum deviances for values of π between 0.5 and 0.7. Thus there is only a small change in predictive ability when increasing π slightly. However, if the prior belief was that the true π was smaller than 0.513, then the choice of π would influence the prediction deviance considerably. It should be noted here that although Figure 6(a) suggests we could estimate π as that which minimizes the deviance, this would be incorrect; in the next section we show that π is not estimable

when fitting a boosted trees model.

The marginal effects on η of all variables follow a pattern similar to that in Figure 6(b). Across different values of π, the shape of the marginal effect is relatively constant, with the magnitude of the effect increasing with π. Unsurprisingly, as π gets smaller the effect size tends to resemble that of the naive model (not shown), as we are assuming that there are few presences in the background data. Thus, in our example, the shapes of the marginal effects are not sensitive to changes in π, but the magnitudes of the effects are.

4. Estimating π

Ideally we would like to be able to estimate the population prevalence π, as well as η, from presence-only data. However, if we make no assumptions about the structure of η (such as in a boosted trees model), then π and η are not identifiable. This means that if we know neither π nor η, then we cannot estimate them accurately. In fact the only way to estimate π is by making unrealistic assumptions about the form of η. In this section we prove conditions for the identifiability of π and η; an illustration of how it may work in practice is given in the Web Appendix.

First we give a heuristic proof that π and η are not identifiable when there is no constraint on the structure of η. Let $f_1(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y=1)$ and $f_0(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y=0)$ be the densities of the covariates $\mathbf{x}$ for the populations of true presences and absences respectively. Then the overall density of $\mathbf{x}$ is $f(\mathbf{x}) = \pi f_1(\mathbf{x}) + (1-\pi) f_0(\mathbf{x})$, and by Bayes rule,

$$\mathbb{P}(y=1|\mathbf{x}) = \frac{\mathbb{P}(y=1)\mathbb{P}(x|y=1)}{\mathbb{P}(x)} = \frac{\pi f_1(\mathbf{x})}{(1-\pi) f_0(\mathbf{x}) + \pi f_1(\mathbf{x})}.$$

The model of interest here is $\text{logit}(\mathbb{P}(y=1|\mathbf{x})) = \eta(\mathbf{x})$, so

$$\eta(\mathbf{x}) = \log\left(\frac{\pi}{1-\pi}\right) + \log\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}\right).$$

Thus η defines, up to a multiplicative constant, the form of the ratio between f_1 and f_0.

The background data provides information about the form of f, as it is a random sample from the population of sites. Similarly we learn about the shape of f_1 from the naive presences. However we have no information about f_0 directly, so if η is unrestricted and π unknown, there is no restriction on the form of f_0. To see this, note that if (π, f_0, f_1, f) is one representation of the densities, then we can write

$$f = \pi f_1 + (1-\pi) f_0 = \pi^* f_1 + (1-\pi^*) f_0^*$$

where $f_0^* = \frac{1}{1-\rho\pi}\left((1-\pi) f_0 + (1-\rho)\pi f_1\right)$ is a density and $\pi^* = \rho\pi$, $0 < \rho < 1$. As f and f_1 remain the same, (π^*, f_0^*, f_1, f) is a family of possible alternate representations. Thus π and f_0 are not uniquely defined without any assumption on the structure of η.

We next prove identifiability of π and η in a conditional logistic regression model, where we assume $\eta(\mathbf{x}) = \mathbf{x}^T\beta$ is linear. The observed likelihood for the naive presences/absences is

$$\begin{aligned} L(\eta|\mathbf{z}, X) &= \prod_i \mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)^{z_i}\mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)^{1-z_i} \\ &= \exp\left(\sum_i z_i \log\left(\frac{\mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)}{\mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)}\right) + \sum_i \log\left(\mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)\right)\right). \end{aligned}$$

From (6), this is the density of a curved exponential family (Lehmann & Casella, 1998) with natural parameter $\zeta(\pi, \eta)$ where

$$\zeta_i = \text{logit}(\mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)) = \log\left(\frac{n_p}{\pi n_u}\frac{e^{\eta(\mathbf{x}_i)}}{1 + e^{\eta(\mathbf{x}_i)}}\right).$$

This family is of full rank if $(\pi, \eta) \neq (\pi', \eta') \iff \zeta \neq \zeta'$, and we note that

$$\zeta(\pi, \eta) = \zeta'(\pi', \eta') \iff \frac{\pi'}{\pi} = \frac{e^{-\eta(\mathbf{x}_i)} + 1}{e^{-\eta'(\mathbf{x}_i)} + 1}. \tag{9}$$

We will assume that X includes a column of 1s, so that $\mathbf{x}_i^T\beta$ includes an intercept term, and also that X is of full rank. Because X is full rank, $\beta \neq \beta' \iff X\beta \neq X\beta'$, and if $X\beta \neq X\beta'$ then $\frac{e^{-\mathbf{x}_i^T\beta}+1}{e^{-\mathbf{x}_i^T\beta'}+1}$ can not be constant for all i. Hence, there does not exist π and π' such that $\zeta = \zeta'$. Similarly if $\pi \neq \pi'$, there is no β and β' such that (9) holds.

A stricter proof of non-identifiability for general η follows directly from this argument. As η is unconstrained, it is possible to choose η such that the equality (9) holds for all $\mathbf{x}_i$, for fixed π,π' and η'. More generally, if we can write $\eta(\mathbf{x}) = h(\mathbf{x})^T\beta$ for some basis functions h of $\mathbf{x}$, such as for GAMs, then π and η are identifiable if and only if $h(\mathbf{x})$ is of full rank. This follows immediately from the argument for the logistic regression model.

An alternate way to approach this modeling problem is by estimating the covariate densities f_0 and f_1 directly. If we are willing to assume parametric forms for these densities then we can usually estimate π. For example, assuming f_0 and f_1 are Gaussian, π could be estimated using a method similar to the EM algorithm for Gaussian mixture estimation (e.g. Hastie *et al.*, 2001, p236). However, this estimate may be severely biased if the assumption is not true. In maxent (Phillips *et al.*, 2006), the covariate density f_1 is modeled directly, as a Gibbs distribution. But there is no assumption about the structure of f_0, either directly or indirectly, and so π is not estimable in this setup.

This analysis strongly contra-indicates estimating π from presence-only data. Although it is theoretically possible to estimate π for the linear logistic regression model, there is an implicit assumption that the log of the tilting function f_1/f_0 is linear in $\mathbf{x}$. This

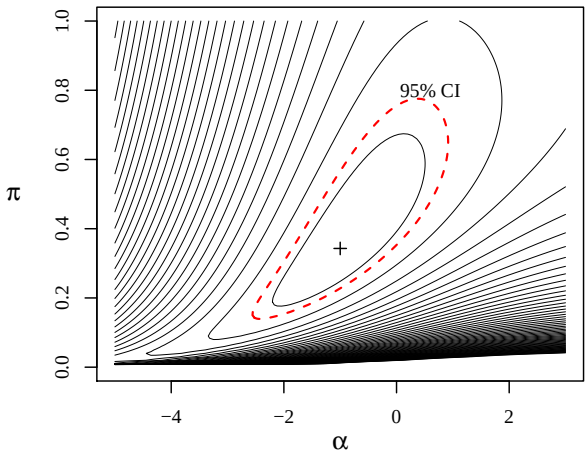


Figure 7. *The likelihood surface, and a 95% confidence interval (dashed line), for π and the intercept α at the true level of β for a simulation of 200 presences and 1000 background data. The true values of $\pi = 0.34$, and $\beta_0 = -1.0$ are indicated on the plot. The generative model is $\eta(x) = -1 + \beta x$, where $\beta = 2$ and the x are i.i.d. stanford normals.*

may be a useful representation of the true model, but not as an assumption upon which π is estimated. In practice estimates of π are highly variable, and correlated with the intercept, even in simulated models where the logistic model is the truth (Figure 7). Where the true densities deviate from the assumed structure, the estimates are highly unstable; Lancaster & Imbens (1996) report failure of convergence of the direct maximization of the observed likelihood in several examples.

5. Conclusions

We have proposed an application of the EM algorithm that provides a flexible method of estimating the underlying presence-absence model from presence-only data. It can work as a wrapper to (almost) any logistic modeling procedures. At each iteration the probabilities of presence are calculated, given the current model, for the background data (E-step) and the case-control logistic model is re-estimated using these new weighted or non-binary outcomes (M-step). For stepwise logistic modeling procedures, such as boosted trees, E-steps may be interleaved within the procedure to save computation time.

This EM model gives approximately unbiased estimates in a simple linear logistic regression simulation study. In comparison, there is considerable shrinkage toward zero in the estimates from the naive logistic regression model, fitted directly to the naive presences and absences. To create a more realistic example, but where we knew the true presences, we generated presence-only data from presence-absence data of eels in

New Zealand rivers. The boosted trees EM model for these data outperform the naive boosted trees model; there is less shrinkage of the marginal effects and the prediction deviance for the presence-absence validation set is smaller for the EM model. Unsurprisingly, however, the EM model still has higher prediction deviance than the boosted trees model fitted using the true presences.

Previous work in estimating the presence-absence model from presence-only data has attempted to simultaneously estimate the population prevalence π. However, we have shown that π is not identifiable when no assumptions are made about the structure of η, such as in boosted trees models. In addition, even when some assumption is made about the structure of η, e.g. for logistic regression models with η linear in the covariates $\mathbf{x}$, the resulting estimate of π is highly variable and heavily dependent on that assumption. Assuming that η is linear in $\mathbf{x}$ is equivalent to assuming that the densities of locations, for presences and absences respectively, are linked by a linear exponential tilt. In the opinion of the authors, although this may be a useful modeling approximation, the estimation of π relies too heavily on this assumption. We recommend obtaining an estimate of π from some other source and using sensitivity analysis to assess the dependence of the results on this estimate.

This application of the EM algorithm provides a flexible way of estimating species distribution from the extensive records of species presence in herbaria and museums, and from radiotelemetry studies. Because of its simplicity, we believe it can be easily adopted by anyone working in this field.

Supplementary Material

The Web Appendix referenced in Section 4 is available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgements

Fish distribution data were taken from the New Zealand Freshwater Fish Database, which is curated by New Zealand's National Institute of Freshwater and Atmospheric Research.

The authors are very grateful to Holger Höfling for identifying the naive likelihood as the density of a curved exponential family. We would also like to thank Rob Tibshirani, Art Owen and the Hastie-Tibshirani Research Group of Stanford University for their helpful feedback on this work.

REFERENCES

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

Dudí?k, M., Schapire, R. and Phillips, S. (2006). Correcting sample selection bias in maximum entropy density estimation. In Weiss, Y., Schölkopf, B. and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 323–330. MIT Press, Cambridge, MA.

Elith, J., Graham, C. H., Anderson, R. P., Dudí?k, M., Ferrier, S., Guisan, A. and Others (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151.

Frair, J. L., Nielsen, S. E., Merrill, E. H., Lele, S. R., Boyce, M. S., andG B Stenhouse, R. H. M. M. and Beyer, H. L. (2004). Removing gps collar bias in habitat selection studies. *Journal of Applied Ecology* **41**, 210–212.

Friedman, J. H. (2001). Greedy function approximation: the gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T., Tibshirani, R. J. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Keating, K. A. and Cherry, S. (2004). Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management* **68**, 774–789.

Lancaster, T. and Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics* **71**, 145–160.

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T. and Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series.* .

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J. and Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology* **50**, 2034–2051.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Chapman & Hall, London, second edition.

Mace, R. D., Waller, J. S., Manley, T. L., Lyon, L. J. and Zuuring, H. (1996). Relationships among grizzly bears, roads and habitat in the Swan Mountains, Montana. *The Journal of Applied Ecology* **33**, 1395–1404.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Pearce, J. L. and Boyce, M. S. (2005). Modeling distribution and abundance with

presence-only data. *Journal of Applied Ecology* **43**, 405–412.

Reese, G. C., Wilson, K. R., Hoeting, J. A. and Flather, C. H. (2005). Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications* **15**, 554–564.

Ridgeway, G. (2004). gbm: Generalized boosted regression models. R package, version 1.3-5. `http://www.i-pensieri.com/gregr/gbm.shtml`.

Walker, P. A. and Cocks, K. D. (1991). HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* **1**, 108–118.

Zaniewski, A. E., Lehmann, A. and Overton, J. M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157**, 261–280.

WEB APPENDIX: ESTIMATION OF π FOR LOGISTIC REGRESSION

In a logistic regression model, where $\eta(\mathbf{x}) = \mathbf{x}^T\beta$ is linear and X is of full rank, we can estimate both π and β. To do this we develop a modified version of the EM procedure developed in the paper, where π is also estimated at every step. As π is unknown we need to use the full likelihood, equation (4) from the main paper, for the naive and true presences $\mathbf{z}$ and $\mathbf{y}$:

$$\begin{aligned} L(\pi, \eta|\mathbf{y}, \mathbf{z}, X) &= \left(\prod_i \mathbb{P}(y_i|s_i = 1, \mathbf{x}_i)\right)\left(\prod_i \mathbb{P}(z_i|y_i, s_i = 1, \mathbf{x}_i)\right) \\ &= L_1(\pi, \eta|\mathbf{y}, \mathbf{z}, X)\ L_2(\pi|\mathbf{y}, \mathbf{z}, X) \end{aligned}$$

Note that L_2 depends on π only and L_1 depends on π and η only through $\eta^* = \eta + \log\left(\frac{n_p+\pi n_u}{\pi n_u}\right)$. Thus we can re-parameterize (π, η) to (π, η^*), so that L_1 depends only on η^*, and we can maximize L_1 and L_2 separately.

L_1 is maximized by fitting a logistic regression model to $\mathbf{y}$, to obtain $\eta^*(\mathbf{x})$. Additionally, the maximum likelihood estimate for π is easily found by maximizing L_2, using the probabilities calculated in Section 2.2 of the main paper:

$$\begin{aligned} L_2(\pi|\mathbf{y}, \mathbf{z}, X) &= \prod_i \mathbb{P}(z_i|y_i, s_i = 1, \mathbf{x}_i) \\ &= \left(\prod_{i\in P} \mathbb{P}(z_i = 1|y_i = 1, s_i = 1, \mathbf{x}_i)\right)\left(\prod_{i\in \mathbf{U}} \mathbb{P}(z_i = 0|y_i = 1, s_i = 1, \mathbf{x}_i)^{y_i}\right) \\ &= \left(\frac{n_p}{n_p + \pi n_u}\right)^{n_p}\left(\frac{\pi n_u}{n_p + \pi n_u}\right)^{\sum_{i\in\mathbf{U}} y_i} \end{aligned}$$

$$\Rightarrow \frac{\partial}{\partial\pi}\log L_1 = -\left(n_p + \sum_{i\in\mathbf{U}} y_i\right)\frac{n_u}{n_p + \pi n_u} + \sum_{i\in\mathbf{U}} y_i\frac{1}{\pi} \ \Rightarrow\ \hat{\pi} = \frac{1}{n_u}\sum_{i\in\mathbf{U}} y_i$$

Thus we have an intuitive estimate for π - the fraction of the background data that are true presences. From $\{\hat{\pi}, \hat{\eta}^*\}$ we can retrieve $\hat{\eta}$ by subtracting $\log(\frac{n_p+\hat{\pi}n_u}{\hat{\pi}n_u})$ from the intercept.

Maximizing L_1 and L_2 forms the M-step of the EM algorithm. Because both $\log L_1$ and $\log L_2$ are linear in the y_i, the expectation step of the EM algorithm is still to replace each y_i with its expectation conditional on the parameter estimates from the previous step. So the EM algorithm for estimating η and π follows as in Figure 1 of the main paper, except that in the maximization step, before adjusting the intercept of $\hat{\eta}^{*(k)}$, we update the estimate of π: $\pi^{(k)} = \frac{1}{n_u}\sum_{i\in\mathbf{U}} \hat{y}_i^{(k-1)}$.

This algorithm appears to provide an ideal solution for estimating π for presence-only data. However, we have already seen in Figure 7 of the main paper that even

when the logistic regression model is the true generative model, estimates of π and the intercept α are highly correlated. In practice the logistic regression model is both unlikely to be the true model and unverifiable. Although logistic regression models are a useful approximation, to estimate π we are placing a greater reliance on that model being the truth. It is safer to have a good estimate of π from another source, and run sensitivity analyses on the results.