A Maximum Likelihood Estimation as a Joint GLM

Recall that maximizing likelihood for the full data set means maximizing

$$\ell(\theta) = \sum_k \ell_{k,\mathrm{PA}}(\alpha_k, \beta_k) + \ell_{k,\mathrm{PO}}(\alpha_k, \beta_k, \gamma_k, \delta), \tag{23}$$

where

$$\ell_{k,\mathrm{PA}}(\alpha_k, \beta_k) = \sum_{i \in I_{\mathrm{PA}}} -y_{ik} \log\left(1 - e^{-\exp\{\alpha_k + \beta_k' x_i\}}\right) + (1 - y_{ik}) \exp\{\alpha_k + \beta_k' x_i\} \tag{24}$$

$$\ell_{k,\mathrm{PO}_k}(\alpha_k, \beta_k, \gamma_k, \delta) \approx \sum_{i \in I_{\mathrm{PO}_k}} (\alpha_k + \beta_k' x_i + \gamma_k + \delta' z_i) - \sum_{i \in I_{\mathrm{BG}}} w_i e^{\alpha_k + \beta_k' x_i + \gamma_k + \delta' z_i} \tag{25}$$

In this section we discuss how to massage (23) into a large GLM in terms of a common set of $m(p+2)+r$ predictors and coefficients. For the moment, we ignore the sum over I_{PO_k} in (25) and deal with the other two sums. The sum in (24) is the log-likelihood for a Bernoulli GLM with complementary log-log link and the sum over I_{BG} in (25) is the log-likelihood for a weighted Poisson GLM with log link.

Note that at each survey site we have m presence-absence observations, one for every species. Similarly, we will introduce one "dummy" response $y_{ik} = 0$ for each species k at each background site i, for $m(n_{\mathrm{PA}} + n_{\mathrm{BG}})$ total observations. For observation ik, introduce auxiliary indicator variables

$$u_{ik_1,k_2} = \begin{cases} 1 & k_1 = k_2 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$v_{ik} = \begin{cases} 1 & i \in I_{\mathrm{BG}} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

The variable u_k allows parameters to vary by species. For example, α_k is the coefficient for u_k and $\beta_{k,j}$ is the coefficient for the interaction $x_j u_k$. The variable v gives us bias terms that apply only to terms in the presence-only likelihood. Thus γ_k is the coefficient for $u_k v$ and δ_j is the coefficient for $z_j v$.

For example, the linear predictor for count or presence-absence for species k at a survey site with predictors x and z is

$$\alpha_k + \beta_k' x_i = \sum_{1 \le h \le m} (\alpha_h u_{ik,h} + \beta_h' x_i u_{ik,h} + \gamma_h u_{ik,h} v_{ik}) + \delta' z_i v_{ik}, \tag{28}$$

using $v = 0$ because we are predicting for presence-absence data.

To check the proportional bias assumption for variable z_j — that is, to check the assumption that δ_j should be the same for every species — we can augment the model with interactions $z_{j*k} = u_k z_j$ for each k, and test the hypothesis that each of those variables has no effect on the regression.

Let X_{PA} denote the $n_{\mathrm{PA}} \times p$ matrix with all x variables for all the survey sites, and let X_{BG} and Z_{BG} denote all the x and z variables for all the background sites. Then if

$$X = \begin{pmatrix} 1 & X_{\mathrm{PA}} & 0 \\ 1 & X_{\mathrm{BG}} & 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 \\ Z_{\mathrm{BG}} \end{pmatrix}, \tag{29}$$

our likelihood is a large weighted GLM with $m(n_{\mathrm{PA}} + n_{\mathrm{BG}})$ observations and overall design matrix

$$\mathbb{X} = \begin{pmatrix} X & 0 & \cdots & 0 & Z \\ 0 & X & \cdots & 0 & Z \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & X & Z \end{pmatrix}, \quad \text{and coefficients} \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \\ \delta \end{pmatrix}. \tag{30}$$

The weights are w_i for rows corresponding to background site i, and 1 for presence-absence sites. Note that the response family and link function are different for different rows.

Turning to the sums over I_{PO_k} in (25), note that they are linear in the coefficients, so all k sums can be combined to obtain a single linear term of the form $\theta' M$. All the parameters may be estimated simultaneously via a slight modification of iterative reweighted least squares that takes into account the m linear terms.

A.1 Iterative Reweighted Least Squares Using Block Structure

Let $n = n_{\mathrm{PA}} + n_{\mathrm{BG}}$. $\mathbb{X}$ has mn rows and $m(p+2)+r$ columns. In principle, we could form the matrix $\mathbb{X}$ and use standard GLM software to fit the model, but we would pay a very high computational price for estimating multiple species at a time.

The main computational bottleneck in each iteration is solving a large weighted linear least-squares problem with mn equations (one per species per site) and $m(p+2)+r$ unknowns. The update for step t requires solving a weighted linear least-squares problem with row weights $W^{(t)} = \mathrm{diag}\left(w^{(t)}\right)$ and working responses $u^{(t)}$:

$$\min_{\theta} \left\| W^{(t)} \left(\mathbb{X}\theta - u^{(t)} \right) \right\|_2^2. \tag{31}$$

Solving a completely general problem of the form (31) would require $\mathcal{O}(m^3 np^2 + mnr^2)$ floating point operations. Fortunately, we can store and compute much more cheaply if we exploit the special block structure of $\mathbb{X}$.

Our computational scheme relies heavily on the following well-known and highly useful lemma:

Lemma 1 (Partitioned Least Squares)**.** *Consider the least-squares problem*

$$\min_{v} \left\| (A \;\; B) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - c \right\|_2^2. \tag{32}$$

Let $B_{\cdot A}$ represent the matrix B with each column orthogonalized with respect to the column space of A. Then for v^* solving (32) *we have*

$$B'_{\cdot A} B_{\cdot A} v_2^* = B'_{\cdot A} c = B'_{\cdot A} c_{\cdot A}. \tag{33}$$

That is, the least-squares coefficients for B may be obtained by first regressing the columns of B on A, then regressing c on the residuals.

Proof. Let M be least-squares coefficients for regression of B on A; that is,

$$B = AM + B_{\cdot A} \tag{34}$$

Then, (32) is equivalent to the least-squares problem

$$\min_{\overline{v}} \left\| \begin{pmatrix} A & B_{\cdot A} \end{pmatrix} \begin{pmatrix} \overline{v}_1 \\ \overline{v}_2 \end{pmatrix} - c \right\|_2^2. \tag{35}$$

To see why, note that

$$A\overline{v}_1 + B_{\cdot A}\overline{v}_2 = A\left(\overline{v}_1 - Mv_2\right) + B\overline{v}_2 \tag{36}$$

so solutions to (32) and (35) are in direct correspondence with one another, with $v_2 = \overline{v}_2$.

Moreover, because the two blocks in (35) are orthogonal to each other, we can solve the problem by separately regressing c on A and on $B_{\cdot A}$ to obtain $\overline{v}_1^*$ and $v_2^* = \overline{v}_2^*$. □

Our proof implies further that having obtained M and v_2^*, we can compute $v_1^* = \overline{v}_1^* - Mv_2^*$.

A.2 Least Squares with Block Structure

Suppressing the t superscript, we need to solve a least squares problem with design matrix $W\mathbb{X}$ and response vector u. Writing

$$W = \begin{pmatrix} W_1 & & \\ & \ddots & \\ & & W_m \end{pmatrix}, \tag{37}$$

we have

$$W\mathbb{X} = \begin{pmatrix} W_1X & & W_1Z \\ & \ddots & \vdots \\ & W_mX & W_mZ \end{pmatrix} = \begin{pmatrix} X_1 & & Z_1 \\ & \ddots & \vdots \\ & X_m & Z_m \end{pmatrix}. \tag{38}$$

Let $\theta_1^*, \ldots, \theta_{m+1}^*$ be the blocks of least-squares coefficients corresponding to the column blocks in (38). Writing $W\mathbb{X} = (\mathcal{X}\ \mathcal{Z})$, Lemma 1 means that given $\widetilde{\mathcal{Z}} = \mathcal{Z}_{\cdot\mathcal{X}}$, we can efficiently solve for the coefficients θ_{m+1} by solving the $r \times r$ system

$$\widetilde{\mathcal{Z}}'\widetilde{\mathcal{Z}}\theta_{m+1}^* = \widetilde{\mathcal{Z}}'u \tag{39}$$

Because $\mathcal{X}$ is block diagonal, the kth row block of $\widetilde{\mathcal{Z}}$ is $\widetilde{Z}_k = Z_{k\cdot X_k}$; that is, orthogonalizing $\mathcal{Z}$ with respect to $\mathcal{X}$ is equivalent to orthogonalizing each Z_k independently with respect to the corresponding X_k. After computing a single QR decomposition of X_k, we compute and store the least-squares coefficients $\overline{\theta}_k$ and Γ_k from regressing u_k and Z_k on X_k. Having done this we can also compute the residuals $\widetilde{Z}_k$ cheaply.

To obtain θ_{m+1}^* in the end, we need only keep a running tally of the quantities appearing in (39),

$$\widetilde{\mathcal{Z}}'\widetilde{\mathcal{Z}} = \sum_k \widetilde{Z}_k'\widetilde{Z}_k, \quad \text{and} \quad \widetilde{\mathcal{Z}}'u = \sum_k \widetilde{Z}_k'u_k, \tag{40}$$

and solving (39) gives θ_{m+1}^*. Now, per Lemma (32), we can reconstruct all of θ^* if we retain the least-squares coefficients of u and Z_k on X_k at every step. Algorithm 1 gives the full details of the procedure.

Most of the computational will typically be spent computing the QR decompositions of the blocks X_k. Each QR decomposition requires $\mathcal{O}(np^2)$ operations, so that $\mathcal{O}(mnp^2)$ total

operations are required for this step. Computing $\widetilde{\mathcal{Z}}'\widetilde{\mathcal{Z}}$ requires $\mathcal{O}(mnr^2)$ operations. Thus our method requires $\mathcal{O}(mn(p^2+r^2))$ operations, compared to $\mathcal{O}(m^3np^2+mnr^2)$ required for the naive method. For $m=36$ species with $p\approx r$, for example, our method does roughly 650 times less work than the naive approach.

Our method is also lightweight with respect to its storage costs. After one block's computation is completed in the first for loop of Algorithm 1, we do not need to store u_k, Z_k, X_k, or its QR decomposition. We need only store the $p(r+1)$ least-squares coefficients from each step.

Algorithm 1: Efficient Least-Squares Using Block Structure of $W\mathbb{X}$

$A \leftarrow 0_{r\times r}$;
$b \leftarrow 0_r$;
for $k = 1,\ldots,m$ **do**
  Compute QR decomposition of X_k;
  Regress Z_k on X_k to obtain $Z_k = X_k\Gamma_k + \widetilde{Z}_k$;
  $A \leftarrow A + \widetilde{Z}_k'\widetilde{Z}_k$;
  Regress u_k on X_k to obtain $u_k = X_k\overline{\theta}_k + \tilde{u}_k$;
  $b \leftarrow b + \widetilde{Z}_k'\tilde{u}_k$;
end
Solve $A\theta^*_{m+1} = b$ for θ^*_{m+1};
for $k = 1,\ldots,m$ **do**
  $\theta^*_k \leftarrow \overline{\theta}_k - \Gamma_k\theta^*_{m+1}$;
end