Mark W. Donoghoe* and Ian C. Marschner

Flexible Regression Models for Rate Differences, Risk Differences and Relative Risks

Abstract: Generalized additive models (GAMs) based on the binomial and Poisson distributions can be used to provide flexible semi-parametric modelling of binary and count outcomes. When used with the canonical link function, these GAMs provide semi-parametrically adjusted odds ratios and rate ratios. For adjustment of other effect measures, including rate differences, risk differences and relative risks, non-canonical link functions must be used together with a constrained parameter space. However, the algorithms used to fit these models typically rely on a form of the iteratively reweighted least squares algorithm, which can be numerically unstable when a constrained non-canonical model is used. We describe an application of a combinatorial EM algorithm to fit identity link Poisson, identity link binomial and log link binomial GAMs in order to estimate semi-parametrically adjusted rate differences, risk differences and relative risks. Using smooth regression functions based on B-splines, the method provides stable convergence to the maximum likelihood estimates, and it ensures that the estimates always remain within the parameter space. It is also straightforward to apply a monotonicity constraint to the smooth regression functions. We illustrate the method using data from a clinical trial in heart attack patients.

Keywords: generalized additive models, B-splines, semi-parametric regression, risk models

DOI 10.1515/ijb-2014-0044

1 Introduction

Binary and count response data are commonly encountered in biostatistical settings, and generalized linear models (GLMs) are used to estimate the adjusted effects of several covariates on the risk or rate of these outcomes. The link function in the GLM determines the scale on which the effect measure is expressed. Using the canonical link with a binomial model for binary data gives adjusted odds ratios, and with a Poisson model for count data gives adjusted rate ratios [1].

In biostatistics, other effect measures such as adjusted rate differences, risk differences or relative risks are often of interest, meaning that a non-canonical link function must be used. Under such models, constraints on the parameter space are required to ensure that fitted rates are non-negative and fitted risks lie within $[0, 1]$. However, the fitting procedures for GLMs in standard statistical software typically rely on a form of the iteratively reweighted least squares algorithm, which can be numerically unstable and fail to converge when parameter space constraints are present [2]. General step-size optimization approaches can improve stability when estimates are close to the boundary of the parameter space, but these are not guaranteed to converge in all situations.

Marschner [3], Donoghoe and Marschner [4] and Marschner and Gillett [5] have described stable methods for finding the maximum likelihood estimate (MLE) of Poisson GLMs with an identity link, and binomial GLMs with identity and log links. The methods allow estimation of adjusted rate differences, risk differences and relative risks respectively, avoiding convergence problems. All of these are applications of the combinatorial expectation–maximization (CEM) algorithm presented by Marschner [6].

*Corresponding author: Mark W. Donoghoe, Department of Statistics, Macquarie University, New South Wales 2109, Australia; NHMRC Clinical Trials Centre, University of Sydney, New South Wales 2006, Australia, E-mail: mark.donoghoe@mq.edu.au

Ian C. Marschner, Department of Statistics, Macquarie University, New South Wales 2109, Australia; NHMRC Clinical Trials Centre, University of Sydney, New South Wales 2006, Australia, E-mail: ian.marschner@mq.edu.au

With these models, however, the functional form of any continuous covariates must be specified. Generalized additive models (GAMs) are an extension of GLMs that allow for extra flexibility through the inclusion of semi-parametric terms [7,8]. This can potentially lead to a better model fit, or help to identify a more parsimonious model for the outcome. Model-fitting with GAMs often uses similar algorithms as those used for GLMs, and hence is subject to similar convergence issues, particularly with non-standard link functions. In fact, in some GAM packages such as `PROC GAM` in SAS [9], only the canonical link is permitted. In other packages, such as the R packages discussed in Section 6, non-canonical links are permitted but can be numerically unstable.

In this paper, we extend the existing CEM algorithms for GLMs to GAMs, by the addition of smooth semi-parametric functions based on B-splines [10,11]. We begin by defining the general GAM in Section 2, and in Section 3 we explore the properties of the B-spline basis functions. In Section 4 we explain the method for finding the MLE of each of these models, as well as how to apply an optional monotonicity constraint to the smooth functions. In Section 5, we demonstrate our methods by applying them to data from a clinical trial in heart attack patients. In Section 6, we summarize other popular methods for fitting GAMs and their performance in this dataset.

2 Model specification

Consider a sample of independent random variables $\{Y_1, \dots, Y_n\}$ with either

$$Y_i \sim \text{Bin}(N_i, \lambda_i) \quad \text{or} \quad Y_i \sim \text{Poisson}(N_i \lambda_i),$$

so that $\mathbb{E}(Y_i) = N_i \lambda_i$ for some fixed known N_i . The quantity λ_i is interpreted as a standardized mean which will be a probability (risk) for the binomial model and a rate for the Poisson model.

In a GAM, the standardized mean λ_i is related to a linear combination of covariates through the link function

$$g(\lambda_i) = \Lambda(\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i; \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^A \alpha_a(u_{ia}) + \sum_{b=1}^B \beta_b v_{ib} + \sum_{c=1}^C f_c(w_{ic}). \quad (1)$$

Here, $\mathbf{u}_i = (u_{i1}, \dots, u_{iA})$ are categorical covariates, where, without loss of generality, each u_{ia} takes a discrete value in $\{1, \dots, k_a\}$. Linear continuous covariates are denoted by $\mathbf{v}_i = (v_{i1}, \dots, v_{iB})$, with each v_{ib} allowed to take any value in the range $[v_b^{(0)}, v_b^{(1)}]$, where $v_b^{(0)} = \min_i \{v_{ib}\}$ and $v_b^{(1)} = \max_i \{v_{ib}\}$. The flexible part of the model is included through the unspecified non-parametric functions $f_1, \dots, f_C$, which take as input the continuous covariates $\mathbf{w}_i = (w_{i1}, \dots, w_{iC})$, and have domain $w_{ic} \in [w_c^{(0)}, w_c^{(1)}]$.

In order to impose a finite dimensional structure on the problem, the MLE $\hat{f}_c$ of the unknown f_c is restricted to be in the space defined by a specified set of basis functions $B_{c1}, \dots, B_{cD_c}$. That is, each $\hat{f}_c$ can be expressed as

$$\hat{f}_c(w) = \sum_{d=1}^{D_c} \gamma_{cd} B_{cd}(w), \quad (2)$$

a restriction we denote by $\hat{f}_c \in \mathcal{B}_c$. Thus, the problem becomes one of finding the MLE of the parameter vector $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\alpha} = (\alpha_1(1), \dots, \alpha_A(k_A))$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_B)$ and $\boldsymbol{\gamma} = (\gamma_{11}, \dots, \gamma_{CD_c})$.

One possible choice of basis functions is a sequence of indicator functions that serve as the increments between successive unique observed values of the covariate. That is, if $z_c(0) < \dots < z_c(D_c)$ are the $D_c + 1$ ordered unique values of w_{ic} , we can define D_c basis functions

$$B_{cd}(w) = \mathbf{1}\{w \geq z_c(d)\}, \quad d = 1, \dots, D_c.$$

The resulting $\hat{f}_c$ is a step function, and each γ_{cd} represents the change in $g(\lambda)$ associated with an increase in the covariate from $z_c(d-1)$ to $z_c(d)$. If we constrain these increments to be strictly non-negative, this is the semi-parametric isotonic model described by Marschner [3], Donoghoe and Marschner [4] and Marschner and Gillett [5].

Under the step function model, a large number of degrees of freedom are sacrificed in order to ensure that the semi-parametric function estimate fits closely to the observed data. Here we will instead focus on a smooth semi-parametric regression technique, in which the number of parameters that need to be estimated is reduced and $\hat{f}_c$ is smooth.

This could be achieved with any of a wide range of basis functions [8]. We will use the polynomial B-splines, as they are highly flexible and their properties allow us to easily integrate them into the existing CEM algorithms, as discussed in the next section.

3 B-splines

3.1 Definition and properties

The B-splines are a series of basis functions for polynomial splines, defined by a grid of q_c fixed turning points $\xi_{c1} < \dots < \xi_{cq_c}$, where $\xi_{c1} = w_c^{(0)}$ and $\xi_{cq_c} = w_c^{(1)}$ are positioned at the extremes of the domain. The grid is expanded to form a sequence τ_c of knots that determine the degree and continuity of the resulting curve [12].

We will restrict our focus to the case of B-splines of order 3, where each basis function is made up of a series of quadratic curves between each pair of turning points, with adjacent curves constrained to have equal gradient at their boundaries. The sequence of $q_c + 4$ knots is defined such that there are three knots at the lower boundary, three knots at the upper boundary and one at each of the internal turning points. That is,

$$\begin{aligned}\tau_{c1} &= \tau_{c2} = \tau_{c3} = \xi_{c1} \\ \tau_{cd} &= \xi_{c(d-2)} \quad \text{for } d = 4, \dots, q_c + 1 \\ \tau_{c(q_c+2)} &= \tau_{c(q_c+3)} = \tau_{c(q_c+4)} = \xi_{cq_c}.\end{aligned}$$

Given a knot sequence, the $D_c = q_c + 1$ B-splines can be calculated recursively [10]. We use $B_{cd}(w)$ to represent the d th basis function $B_d(w; \tau_c)$ on knot sequence τ_c , and $\mathcal{B}_c = \mathcal{B}(\tau_c)$ to denote the associated function space for $\hat{f}_c$.

The B-splines are normalized such that

$$\sum_{d=1}^{D_c} B_{cd}(w) = 1 \quad (3)$$

for all w , meaning that a constraint must be applied to the parameters for each c to ensure that they are identifiable. We do this by setting $\gamma_{ct_c} = 0$ for some choice of $t_c \in \{1, \dots, D_c\}$.

Each basis function $B_{cd}(w)$ is positive for all $w \in (\tau_{cd}, \tau_{c(d+3)})$ and zero elsewhere, taking its maximum value for some $w \in (\tau_{c(d+1)}, \tau_{c(d+2)})$. This means that each parameter γ_{cd} has only local influence on the smooth function $\hat{f}_c$.

As an illustration we present a graphical representation in Figure 1. The four B-spline basis functions of order 3 with a single internal turning point, and hence 7 knots, are shown in Figure 1(A). Figure 1(C) shows the curve $\hat{f}$ that results from taking a particular linear combination of these basis functions:

$$\hat{f}(w) = 0.009B_1(w) + 0.009B_2(w) + 0.055B_3(w) + 0.235B_4(w), \quad (4)$$

which can be equivalently expressed as

$$\hat{f}(w) = 0.009 + 0.000B_2(w) + 0.045B_3(w) + 0.226B_4(w)$$

by using eq. (3). Figure 1(B) and (D) illustrates the use of B-splines to restrict $\hat{f}$ to be monotonic, and will be discussed in Section 3.3.

Figure 1: (A) B-spline basis functions and (C) example resulting curve with one internal knot, and (B) the equivalent monotonic B-spline basis functions and (D) resulting curve based on the same coefficients.

3.2 Model constraints

The models considered in this paper have constraints on their parameter spaces due to the restricted range of the response variable: rates must be non-negative and risks must lie in $[0, 1]$. These constraints are often the source of instability in standard fitting algorithms, particularly when the MLE is close to the boundary of the parameter space [2].

The CEM algorithm is ideal for fitting these models, as it applies the parameter space constraints while guaranteeing stable convergence to the MLE. A crucial step in the definition of a CEM algorithm is that the parameter space is partitioned into a sequence of subspaces, each of which corresponds to a particular set of constraints on individual parameters. The properties of the B-splines as discussed in Section 3.1 allow us to extend the existing methods to include these semi-parametric functions.

As an example, we restrict our attention to $\hat{f}_c \in \mathcal{B}_c^+$, the space of strictly non-negative curves in $\mathcal{B}_c$. This can be done without loss of generality due to eq. (3), such that the range of $\hat{f}_c$ is determined by the intercept α_0 , and its shape by the B-spline coefficients.

The function space $\mathcal{B}_c^+$ can be partitioned into subspaces defined by the index of the smallest coefficient. That is, if we define

$$\mathcal{B}_c^+(t_c) = \left\{ f \in \mathcal{B}_c^+ : \gamma_{ct_c} = \min_d \{\gamma_{cd}\} \right\}, \quad (5)$$

it is easy to see that

$$\mathcal{B}_c^+ = \bigcup_{t_c=1}^{D_c} \mathcal{B}_c^+(t_c). \quad (6)$$

For a particular identifiability constraint $\gamma_{ct_c} = 0$, if the remaining coefficients are restricted to be non-negative, the resulting $\hat{f}_c$ will be a strictly non-negative curve. Furthermore, γ_{ct_c} will be the smallest of the coefficients, that is, $\hat{f}_c \in \mathcal{B}_c^+(t_c)$.

One characterization of the constrained function space is that any curve $\hat{f}_c(w) \in \mathcal{B}_c^+(t_c)$ will have a local minimum for $w \in [\tau_{c(t_c+1)}, \tau_{c(t_c+2)}]$. The special case of $t_c = 1$ corresponds to the family of non-negative curves that take their minimum at the lower limit of the domain, $w_c^{(0)}$, and likewise $t_c = D_c$ will constrain the curves to take their minimum at $w_c^{(1)}$. This is demonstrated with an example in Section 4.2.

An analogous result can be achieved by fixing $\gamma_{ct_c} = 0$ and restricting the remaining parameters to be non-positive. Then $\hat{f}_c(w)$ will be a non-positive curve with γ_{ct_c} being the largest coefficient, a function space we denote by $\mathcal{B}_c^-(t_c)$. Similarly to the non-negative case discussed above, $\mathcal{B}_c^-(t_c)$ can be characterized by its restriction on the location of local maxima.

A useful property of the existing CEM algorithms is that they can easily accommodate non-negativity or non-positivity constraints on individual coefficients. By applying such constraints, we can find the MLE of the GAM under the restriction that each $\hat{f}_c \in \mathcal{B}_c^+(t_c)$ or $\mathcal{B}_c^-(t_c)$, and by repeating this process for all possible choices of the identifiability constraint, we can find the overall MLE. This process is explained in more detail for the specific models of interest in Section 4.

3.3 Monotonic B-splines

In some contexts, it may be sensible to restrict $\hat{f}_c(w)$ to increase with increasing w . A sufficient condition for $\hat{f}_c$ to be monotonically non-decreasing is that the coefficients of the B-splines are themselves strictly non-decreasing [13], that is, $\gamma_{c1} \leq \dots \leq \gamma_{cD_c}$.

Assuming non-negative coefficients, the identifiability constraint $\gamma_{c1} = 0$ ensures that $\hat{f}_c(w_c^{(0)}) = 0$ is the minimum of the smooth function. We can then introduce $D_c - 1$ new parameters that represent the increments between successive B-spline coefficients:

$$\delta_{cd} = \gamma_{cd} - \gamma_{c(d-1)},$$

for $d = 2, \dots, D_c$. Restricting these increments to be non-negative will then ensure that the original coefficients are monotonically non-decreasing, as desired.

This can be achieved in the same manner as the unrestricted case by re-expressing eq. (2) as

$$\begin{aligned} \hat{f}_c(w) &= \sum_{d=2}^{D_c} \left[\sum_{e=2}^d \delta_{ce} \right] B_{cd}(w) \\ &= \sum_{d=2}^{D_c} \delta_{cd} \left[\sum_{e=d}^{D_c} B_{ce}(w) \right] \\ &= \sum_{d=2}^{D_c} \delta_{cd} B_{cd}^{m+}(w). \end{aligned}$$

Here, B_{cd}^{m+} , $d = 2, \dots, D_c$, denotes a new series of non-negative monotonic basis functions, which we will refer to as monotonic B-splines. In fact, for the order 3 B-splines used here, these new basis functions are equivalent to the piecewise quadratic integrated splines (I-splines) introduced by Ramsay [12] and used by Tutz and Leitenstorfer [14] for generalized smooth monotonic regression.

Continuing with the illustration provided in Section 3.1, the monotonic B-spline basis functions associated with the B-spline bases in Figure 1(A) are shown in Figure 1(B). Because the B-spline coefficients in eq. (4) are monotonically non-decreasing, the resulting curve in Figure 1(C) can be expressed in terms of the monotonic B-splines

$$\hat{f}(w) = 0.009 + 0.000B_2^{m+}(w) + 0.045B_3^{m+}(w) + 0.181B_4^{m+}(w),$$

and this is demonstrated in Figure 1(D), where the intercept term is shown as a horizontal line.

For the alternate case in which we wish for the curve $\hat{f}_c$ to be non-positive and monotonically non-decreasing with its maximum value at $w_c^{(1)}$, a similar process applies. We fix $\gamma_{cD_c} = 0$, and define $D_c - 1$ new parameters, $\delta_{cd} = \gamma_{cd} - \gamma_{c(d+1)}$ for $d = 1, \dots, D_c - 1$. The associated monotonic basis functions are defined as

$$B_{cd}^{m-}(w) = \sum_{e=1}^d B_{ce}(w) \quad d = 1, \dots, D_c - 1, \quad (7)$$

and their coefficients δ_{cd} are constrained to be non-positive.

3.4 Knot selection

The number and placement of the turning points will influence the shape of the resulting function by determining the space in which our estimated $\hat{f}_c$ is constrained to lie. Ideally, external information would be used to determine the turning points; however, in many applications this will not be available and we are forced to depend on our data to guide this decision.

In some situations, such as OLS, the positioning of the turning points is crucially important [15], and a large number of knot selection methods have been derived (e.g. [10], pp. 174–196, [16]). Many of these methods can be integrated into the approach we describe, but they often lead to a large increase in the computational burden. In a general setting, Ramsay [12] noted that the shape of the resulting estimate is not particularly sensitive to knot placement.

We apply an adaptation of cardinal splines discussed by Hastie and Tibshirani ([7], p. 24), placing the $q_c - 2$ internal turning points at evenly spaced quantiles of the observed covariate values w_{ic} . This is often the standard approach used with regression splines (e.g. [17]). Alternatively, our implementation in R discussed in Section 6 optionally allows the user to specify their own list of knots, so other knot-placement regimes could be used, such as equally spaced knots, nested knot structures or the approach proposed by Yao and Lee [18], in which knots are placed at the local minima and maxima of an initial estimate of the smooth curve, fitted using a basic knot structure.

Of more importance is the choice of the number of turning points. With too few, we may fail to detect important features of the relationship, but with too many, we are at risk of over-fitting. One common way to resolve this trade-off is to use a sufficiently large number of turning points to broaden the function space for $\hat{f}_c$, but add a penalty term to the likelihood function such that spurious fluctuations in the smooth function are avoided [19].

However, with a penalty term, the CEM algorithm central to these stable methods cannot be used directly. We discuss this further, and propose a possible solution in Section 7. In general, we will use the Akaike information criterion $AIC = 2K - 2\ell$ [20] to choose between models with different numbers of knots, where ℓ is the log-likelihood of the fitted model. This similarly includes a penalty for model complexity, and in the case of unpenalized maximum likelihood estimation, the effective degrees of freedom K is simply the number of estimated parameters in the model [8]. The AIC_c is a bias-corrected version of the AIC for small samples [21], which becomes virtually identical as n increases. Figure 4 illustrates the use of the AIC in determining the optimal number of knots for the example in Section 5.

As a criterion for choosing the optimal smoothing parameter in a wide range of scenarios, the AIC and AIC_c have been shown to be generally superior to other classical approaches such as BIC and GCV in the context of general non-parametric regression by Hurvich, Simonoff and Tsai [22], semi-parametric Cox regression by Malloy, Spiegelman and Eisen [23] and likelihood-based boosting by Leitenstorfer and Tutz [13].

Figure 2 shows the results from a simulation study in which we examined the performance of the AIC value in selecting the optimal number of knots. We simulated 500 datasets of 750 binomial observations, each with true risk functions that were based on B-splines with 1, 2 or 3 internal knots. For each dataset, we found the MLE corresponding to a binomial model in which the continuous variable was included as a linear term or as a B-spline with between 0 and 5 internal knots. Of these, the model with the smallest AIC was selected as the optimal model. This approach selected the correct number of knots in the vast majority of samples, and the mean number of knots selected was close to the true value. As has been observed in other contexts, the BIC was biased towards a low number of knots (oversmoothing), and the GCV criterion tended to undersmooth by choosing a higher number of knots more often (data not shown).

The computational effort required by full cross-validation renders it infeasible in this situation. Nevertheless, the focus of this paper is on providing a method of estimation for a single model, which may be applied within any scheme for determining the optimal number of knots.

Figure 2: Results from 500 simulations of the performance of the AIC for selecting the optimal number of knots. The panels on the left-hand side show the true risk function $p(w)$ with (A) 1, (B) 2 and (C) 3 internal knots. The panels on the right-hand side show a histogram of the model selected by AIC in each simulation, with the bar corresponding to the true model shaded.

4 Method

4.1 CEM algorithm

Each of the methods that we use to fit the models in the subsequent sections is an application of CEM algorithms [6]. A CEM algorithm is a general approach in which we consider a finite family of complete data models, indexed by $t \in \mathcal{T}$, each of which has a parameter space $\Theta(t)$ that is a subset of the parameter space Θ for the model of interest, such that

$$\bigcup_{t \in \mathcal{T}} \Theta(t) = \Theta. \quad (8)$$

The complete data models are defined such that an expectation–maximization (EM) algorithm [24] can be used to find the constrained maximum of the likelihood $\hat{\theta}(t)$, within each $\Theta(t)$. Then, due to eq. (8), the $\hat{\theta}(t)$ that attains the highest likelihood is the MLE $\hat{\theta}$.

When the model of interest is a GLM, the complete data models typically impose some constraint on the individual model parameters, such as non-negativity. In the case of GAMs as described below, we use augmented complete data models, where the effect of such constraints on the spline coefficients is to constrain the shape of the smooth curves $\hat{f}_c$. The overall MLE will then correspond to one of these constrained maxima. We now describe the details for each of the specific models of interest.

4.2 Adjusted rate differences

Adjusted rate differences can be estimated by fitting an identity link Poisson GAM. Specifically, if N_i is the period of time over which Y_i events were observed, and we assume $Y_i \sim \text{Poisson}(N_i \lambda_i)$, then λ_i is the event rate for an individual with covariate vector $(\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i)$.

With link function $g(\lambda) = \lambda$, the absolute rate difference is simply the difference between two linear functions. Keeping all other covariates constant, the adjusted rate difference associated with changing the a th categorical covariate from level u_{ia} to u_{ja} is $\alpha_a(u_{ja}) - \alpha_a(u_{ia})$. Likewise, the adjusted rate difference for a 1-unit increase in the b th linear covariate is β_b .

The Poisson means must be non-negative, and so the parameter space within which the MLE $\hat{\boldsymbol{\theta}}$ must lie is

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\mathbf{u}, \mathbf{v}, \mathbf{w}; \boldsymbol{\theta}) \geq 0, (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\},$$

where $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ denote the covariate spaces defined by the Cartesian products of the observed ranges of covariate values:

$$\mathcal{U} = \prod_{a=1}^A \{1, \dots, k_a\}, \quad \mathcal{V} = \prod_{b=1}^B [v_b^{(0)}, v_b^{(1)}], \quad \mathcal{W} = \prod_{c=1}^C [w_c^{(0)}, w_c^{(1)}]. \quad (9)$$

For $C = 0$, Marschner [3] has described a stable CEM algorithm to find the MLE $\hat{\boldsymbol{\theta}} \in \Theta$. This is achieved by partitioning the parameter space into a sequence of distinct constrained parameter spaces $\Theta'(\mathbf{r}, \mathbf{s})$, in which $(\mathbf{r}, \mathbf{s}) = (r_1, \dots, r_A, s_1, \dots, s_B)$ is the covariate vector associated with the minimum fitted Poisson mean.

For a particular choice of $\mathbf{r}$ and $\mathbf{s}$, a complete data model is defined, consisting of $A + B + 1$ independent latent random variables underlying each observed random variable Y_i . These latent variables have Poisson distributions with means that each depend on just one of the model parameters. Using this separation of parameters, an EM algorithm is then defined to find the constrained MLE $\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{s}) \in \Theta'(\mathbf{r}, \mathbf{s})$.

By considering all possible reference vectors $(\mathbf{r}, \mathbf{s})$ that could result in the minimum fitted mean, and finding the MLE within each corresponding constrained parameter space, we can be sure that one of these constrained MLEs will be the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$.

For $C > 0$, it is straightforward to extend this idea to handle smooth semi-parametric components. The reference vector is expanded to include $\mathbf{t} = (t_1, \dots, t_C)$, where each $t_c \in \{1, \dots, D_c\}$. For a particular choice of $\mathbf{t}$, the identifiability constraint $\gamma_{ct_c} = 0$ is applied for each $c = 1, \dots, C$.

The complete data model is augmented with $\sum_c (D_c - 1)$ latent Poisson random variables $Y_i^{(cd)}$, each having an expected value that depends on one of the remaining parameters, that is,

$$Y_i^{(cd)} \sim \text{Poisson}(N_i \gamma_{cd} B_{cd}(\mathbf{w}_{ic})), \quad d \neq t_c. \quad (10)$$

If the basis function values $B_{cd}(\mathbf{w}_{ic})$ are viewed as additional continuous covariates, this augmented complete data model has the same form as that used by Marschner [3]. Thus the EM algorithm can be applied directly to find its MLE $\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{s}, \mathbf{t})$, noting that due to eq. (10), each γ_{cd} is constrained to be non-negative.

As discussed in Section 3.2, the non-negativity constraints on the γ_{cd} force each estimated $\hat{f}_c$ to belong to $\mathcal{B}_c^+(t_c)$ as defined in eq. (5). The constrained MLE will then belong to the parameter space

$$\Theta(\mathbf{r}, \mathbf{s}, \mathbf{t}) = \Theta(\mathbf{r}, \mathbf{s}) \cap \left\{ \bigcap_{c=1}^C \{\boldsymbol{\theta} : \hat{f}_c \in \mathcal{B}_c^+(t_c)\} \right\},$$

where

$$\Theta(\mathbf{r}, \mathbf{s}) = \{\boldsymbol{\theta} : \Lambda(\mathbf{u}, \mathbf{v}, \mathbf{w}) \geq \Lambda(\mathbf{r}, \mathbf{s}, \mathbf{w}) \geq 0, (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\}.$$

Let $\mathcal{R}$ denote the set of all possible choices for $\mathbf{r}$, $\mathcal{S}$ the set of possible choices for $\mathbf{s}$ and $\mathcal{T}$ the set of possible choices for $\mathbf{t}$. Then due to eq. (6), we have that

$$\Theta = \bigcup_{\mathbf{r} \in \mathcal{R}} \bigcup_{\mathbf{s} \in \mathcal{S}} \bigcup_{\mathbf{t} \in \mathcal{T}} \Theta(\mathbf{r}, \mathbf{s}, \mathbf{t}).$$

We have thus defined a CEM algorithm for finding the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$, which will be the constrained estimate $\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{s}, \mathbf{t})$ associated with the highest likelihood.

This approach can be implemented directly using the existing CEM algorithm for identity link Poisson GLMs. Using this algorithm, it is straightforward to apply non-negativity constraints to the coefficients of some continuous covariates by considering only one of the two possible reference levels for that covariate.

For a particular choice of $\mathbf{t}$, if we apply the CEM algorithm to the categorical and linear covariates as usual, and include the basis function values $B_{cd}(w_{ic})$ (for $d \neq t_c$) as linear covariates with non-negativity constraints, we will find a constrained MLE $\hat{\boldsymbol{\theta}}(\mathbf{t}) \in \Theta(\mathbf{t})$, where

$$\Theta(\mathbf{t}) = \Theta \cap \left\{ \bigcap_{c=1}^C \{ \boldsymbol{\theta} : \hat{f}_c \in \mathcal{B}_c^+(t_c) \} \right\}.$$

Repeating this for all $\mathbf{t} \in \mathcal{T}$ will result in a collection of constrained MLEs $\hat{\boldsymbol{\theta}}(\mathbf{t})$, one of which will be the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$. There are a total of $\prod_{c=1}^C D_c$ elements in $\mathcal{T}$, so by applying the CEM algorithm for each, the maximum total number of applications of the EM algorithm required to find the MLE is

$$\prod_{a=1}^A k_a \times 2^B \times \prod_{c=1}^C D_c.$$

The process can be halted if one such application converges to a point in the interior of the constrained parameter space, as we can be sure that this is the overall MLE.

The process of cycling through the parameter space partition is illustrated in Figure 3 for a Poisson model with $C = 1$ covariate. Here we have simulated 500 observations with covariate values w_i and means $\lambda_i = 50(\alpha_0 + \hat{f}(w_i))$, based on a B-spline with two internal knots

$$\hat{f}(w) = \sum_{d=1}^5 \gamma_d B_d(w),$$

where the usual dependence on c has been suppressed since $C = 1$. By virtue of the constraint (3), one of the γ_d parameters must be set to zero while all others must be non-negative. The specific B-spline used to simulate the data for Figure 3 is

$$\alpha_0 + \hat{f}(w) = 0.1 + 0.2B_1(w_i) + 0.3B_2(w_i) + 0.4B_4(w_i) + 0.3B_5(w_i),$$

which has $\gamma_3 = 0$ and is represented by the red dotted line in each plot. In this example, cycling through the parameter space partition involves computing the constrained MLE for each of the five models corresponding to the constraint $\gamma_d = 0$, $d = 1, \dots, 5$. Thus, we would expect the constrained MLE with $\gamma_3 = 0$ to yield the overall MLE. This is shown in Figure 3(A), while the constrained MLEs obtained by using the other constraints are shown in Figure 3(B)–(E). As expected, the log-likelihood for the curve with $\gamma_3 = 0$ is higher than that for the other constrained MLEs (–1378 compared to –1437 through –1548) so it is the overall MLE.

One point to note is that the non-negativity of the coefficients γ_{cd} is a sufficient, but not necessary, condition for the non-negativity of the estimated $\hat{f}_c$. Thus, the parameter space over which we search for the MLE is not guaranteed to include all $\hat{f}_c \in \mathcal{B}_c^+$. However, in practice it would be highly unlikely that the true relationship we are modelling is exactly a quadratic spline on the selected knot sequence $\boldsymbol{\tau}_c$, and so if we find a MLE that is restricted by this deficiency, we can modify $\boldsymbol{\tau}_c$ in order to achieve a better fit. In testing, we found that simply including an additional knot close to the minimum of the curve will generally resolve this issue.

Figure 3: Illustration of the effect of parameter constraints on the smooth curve in a simulated Poisson dataset. The red dotted line represents the true underlying rate and the grey dotted lines denote the knot locations. The black solid line in each panel is the MLE under the identifiability constraint shown in the heading, with all other parameters constrained to be non-negative. The MLE is obtained with the constraint $\gamma_3 = 0$.

4.3 Adjusted relative risks

With binomial data we have $Y_i \sim \text{Bin}(N_i, \lambda_i)$, where N_i is the number of independent trials over which Y_i binary events were observed, and λ_i is the constant event probability, or risk, at each trial. The adjusted relative risk is the ratio of event probabilities associated with a change in one covariate, keeping the others constant.

With $g(\lambda) = \log(\lambda)$, the risk is $\lambda_i = \exp\{\Lambda(\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i; \boldsymbol{\theta})\}$, and so the relative risk is a ratio of two exponential functions. Keeping the other covariates constant, the adjusted relative risk associated with a change in the a th categorical covariate from u_{ia} to u_{ja} is then $\exp\{\alpha_a(u_{ja}) - \alpha_a(u_{ia})\}$. Similarly, the adjusted relative risk for a 1-unit increase in the b th linear covariate is $\exp(\beta_b)$.

The fitted risks must lie within $[0, 1]$, and so the fitted linear predictors must be strictly non-positive. That is, the parameter space for this model is

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\mathbf{u}, \mathbf{v}, \mathbf{w}; \boldsymbol{\theta}) \leq 0, (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\},$$

where $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ are as defined in eq. (9).

With $C = 0$, Marschner and Gillett [5] provide a CEM algorithm for finding the MLE $\boldsymbol{\theta} \in \Theta$. The approach is similar to that used for the identity link Poisson GLM: for a given reference vector $(\mathbf{r}, \mathbf{s})$, the model is reparameterized and a complete data model is defined, which imposes non-positivity constraints on each of the transformed parameters. An EM algorithm is used to find the MLE $\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{s})$ of the complete data model, which is constrained to lie in the parameter space

$$\Theta'(\mathbf{r}, \mathbf{s}) = \{\boldsymbol{\theta} : \Lambda(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \leq \Lambda(\mathbf{r}, \mathbf{s}; \boldsymbol{\theta}) \leq 0, (\mathbf{u}, \mathbf{v}) \in \mathcal{U} \times \mathcal{V}\}.$$

That is, $(\mathbf{r}, \mathbf{s})$ is the covariate vector associated with the maximum linear predictor, which is constrained to be non-positive.

The union of these constrained parameter spaces over all possible choices of $(\mathbf{r}, \mathbf{s})$ is Θ , and so the constrained MLE with the highest likelihood will be the overall MLE. It is straightforward to apply non-positivity constraints to individual parameters.

Extension to include smooth terms follows an analogous approach to that used in Section 4.2. For a particular $\mathbf{t} = (t_1, \dots, t_C)$, we set $\gamma_{ct_c} = 0$ and apply the CEM algorithm, including the remaining basis function values $B_{cd}(w_{ic})$ as continuous covariates with non-positivity constraints. Then each $\hat{f}_c \in \mathcal{B}_c^-(t_c)$, the space of non-positive B-splines that have their shape constrained by the choice of γ_{ct_c} as the largest coefficient. The resulting estimate $\hat{\boldsymbol{\theta}}(\mathbf{t})$ is the constrained MLE over the parameter space

$$\Theta(\mathbf{t}) = \Theta \cap \left\{ \bigcap_{c=1}^C \{ \boldsymbol{\theta} : \hat{f}_c \in \mathcal{B}_c^-(t_c) \} \right\}.$$

By considering each possible $\mathbf{t} \in \mathcal{T}$, and hence all possible locations for the maximum of each $\hat{f}_c$, we find a collection of constrained MLEs $\hat{\boldsymbol{\theta}}(\mathbf{t})$, and the one with the highest likelihood is then the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$.

Again, the sequence may stop early if a maximum in the interior of the parameter space is identified. Furthermore, the same caution also applies here as it did for the identity link Poisson model: non-positivity of the coefficients is only a sufficient condition for non-positivity of the linear predictor, so we are not searching the entire parameter space where $\hat{f}_c \in \mathcal{B}_c^-$. As before, this may be remedied by adjusting the choice of $\boldsymbol{\tau}_c$ if necessary.

4.4 Adjusted risk differences

Adjusted risk differences can be estimated using a binomial GAM with identity link, that is, $Y_i \sim \text{Bin}(N_i, \lambda_i)$ and $g(\lambda) = \lambda$. The adjusted risk difference associated with a change in the a th categorical covariate from u_{ia} to u_{ja} is $\alpha_a(u_{ja}) - \alpha_a(u_{ia})$. Similarly, β_b represents the adjusted risk difference for a 1-unit increase in the b th linear covariate.

Again we require that the probabilities λ lie within $[0, 1]$, but now the parameter space Θ simultaneously imposes both lower and upper boundaries on the linear predictors, that is,

$$\Theta = \{ \boldsymbol{\theta} : 0 \leq \Lambda(\mathbf{u}, \mathbf{v}, \mathbf{w}; \boldsymbol{\theta}) \leq 1, (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W} \}.$$

With $C = 0$, Donoghoe and Marschner [4] have presented an approach for finding the MLE $\hat{\boldsymbol{\theta}} \in \Theta$. It exploits the multinomial–Poisson transformation [25] to convert the model into an equivalent identity link Poisson fit, and the CEM algorithm of Marschner [3] can then be applied to the transformed data to find the MLE. As with the other CEM algorithms, non-negativity constraints can be applied to some of the parameters, which is achieved by imposing such constraints when employing the identity link Poisson CEM algorithm.

However, the inclusion of the semi-parametric terms cannot proceed in exactly the same way as in Sections 4.2 and 4.3. The covariate space considered for continuous covariates entered into the identity link binomial CEM algorithm is the Cartesian product of the observed ranges of those covariates, that is, $\mathcal{V}$ in eq. (9). So if we include the observed B-spline values $B_{cd}(w_{ic})$ as continuous covariates, the covariate space will include vectors in which more than three of the basis functions for a given c are non-zero. By definition, this cannot correspond to the B-spline values for any w . Since the parameter space constraints are imposed for all points in the covariate space, using a larger covariate space than is necessary means that the parameter space is overly restrictive, and may not include the overall MLE.

Instead we must use a slightly different approach, based on the ordering of the B-spline coefficients. We begin by choosing $\mathbf{t} = (\mathbf{t}_1, \dots, \mathbf{t}_C)$, where each $\mathbf{t}_c = (t_{c1}, \dots, t_{cD_c})$ is now a vector containing some permutation of $\{1, \dots, D_c\}$.

Recall that if we define new basis functions B_{cd}^{m+} as described in Section 3.3, non-negativity constraints on the coefficients δ_{cd} will impose monotonicity on the coefficients of the original basis functions, that is, $\gamma_{c1} \leq \dots \leq \gamma_{cD_c}$. Similarly, for a particular permutation $\mathbf{t}_c$, we can define

$$B_{cd}^{(\mathbf{t}_c)}(w) = \sum_{e=d}^{D_c} B_{ct_{ce}}(w), \quad d = 2, \dots, D_c,$$

and non-negativity constraints on the associated coefficients will impose an order restriction on the original coefficients, that is, $0 = \gamma_{ct_{c1}} \leq \dots \leq \gamma_{ct_{cD_c}}$.

We denote by $\mathcal{B}_c^{(\mathbf{t}_c)+}$ the subspace of $\mathcal{B}_c^+$ in which the coefficients are ordered in this way, and note that the overall MLE must correspond to one such permutation. If for a particular $\mathbf{t}$ we enter the new basis function values $B_{cd}^{(\mathbf{t}_c)}(w_{ic})$ as continuous covariates into the identity link binomial CEM algorithm, and impose non-negativity constraints on their coefficients, the resulting estimate $\hat{\boldsymbol{\theta}}(\mathbf{t})$ will be the MLE for the parameter space

$$\Theta(\mathbf{t}) = \Theta \cap \left\{ \bigcap_{c=1}^C \{ \boldsymbol{\theta} : \hat{f}_c \in \mathcal{B}_c^{(\mathbf{t}_c)+} \} \right\}.$$

There are $D_c!$ possible choices for each $\mathbf{t}_c$, and hence $\prod_{c=1}^C D_c!$ possible choices for the vector $\mathbf{t}$. In order to find the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$, we find the constrained MLE $\hat{\boldsymbol{\theta}}(\mathbf{t})$ for each possible choice of $\mathbf{t}$, and choose the one with the highest likelihood.

As with the other models, we may stop early if we find a constrained MLE in the interior of the parameter space, and there are various strategies to identify the parameterizations that are more likely to contain this MLE in order to potentially reduce computing time [6].

4.5 Monotonic smooth regression

Imposing a monotonicity restriction on one or more of the smooth curves is straightforward for all of the models examined so far. As discussed in Section 3.3, a sufficient condition for $\hat{f}_c$ to be monotonically non-decreasing is that the B-spline coefficients themselves are monotonically non-decreasing, that is, $\gamma_{c1} \leq \dots \leq \gamma_{cD_c}$. Because only one possible choice for the value t_c or the vector $\mathbf{t}_c$ needs to be considered, adding a smooth monotonic covariate to a model requires no additional applications of the EM algorithm.

For the identity link binomial model in Section 4.4, applying this constraint is trivial. In iterating through the possible choices of $\mathbf{t}$, we need only to consider $\mathbf{t}_c = (1, \dots, D_c)$ rather than all possible permutations, giving $0 = \gamma_{c1} \leq \dots \leq \gamma_{cD_c}$.

For the other models, we employ the methods discussed in Section 3.3. In an identity link Poisson model, we replace the B-spline function values $B_{cd}(w_{ic})$ by their monotonic B-spline equivalents, $B_{cd}^{m+}(w_{ic})$, and enter these into the CEM algorithm as continuous covariates with non-negativity constraints on the associated coefficients δ_{cd} . The resulting estimate will have $0 = \gamma_{c1} \leq \dots \leq \gamma_{cD_c}$, as desired.

For the log link binomial model we proceed similarly, replacing $B_{cd}(w_{ic})$ by $B_{cd}^{m-}(w_{ic})$ as defined in eq. (7). These are entered into the CEM algorithm as continuous covariates with non-positivity constraints, ensuring $\gamma_{c1} \leq \dots \leq \gamma_{cD_c} = 0$.

It is important to note that the monotonicity of the coefficients is only a sufficient condition for the monotonicity of the resulting smooth function. Hence the function space over which we search for the MLE $\hat{f}_c$ is only a subset of the space of monotonic functions on a B-spline basis. However, when Tutz and Leitenstorfer [14] used boosting techniques to impose constraints on the monotonicity of the function rather than only the coefficients, they found “much higher computational costs without much effect on performance”.

5 Application

The ASSENT-2 study [26] was a randomized clinical trial designed to assess the safety and efficacy of tenecteplase versus alteplase in 16,949 patients treated within 6 hours of an acute myocardial infarction (MI). The primary outcome was 30-day mortality after randomization, and the primary analysis of this outcome showed that the two treatments were equivalent.

To demonstrate our method, we undertake a risk factor analysis, in which the risk of death is modelled in terms of a semi-parametric age effect and three categorical covariates: MI severity (Killip class I, II or III/IV), treatment delay (< 2 , $2-4$, > 4 hours) and geographic region (Western countries, Latin America or Eastern Europe). In view of the natural relationship between age and death, the age-specific risk is constrained to be monotonically non-decreasing. Figure 4 shows the AIC for models with different numbers of knots, using both log and identity link functions.

Figure 4: AIC of the fitted binomial model in the ASSENT-2 study with identity (black, solid line) and log (red, dotted line) links, for a varying number of parameters associated with the age covariate.

Figure 4 shows that the inclusion of a linear age term (1 parameter) improves the AIC considerably for both the log link and identity link models. In the risk difference model, allowing the age term to be quadratic (2 parameters) and semi-parametric with one internal knot (3 parameters) further reduces the AIC substantially, but beyond this, the penalty of additional parameters overrides the small improvements in fit. In the relative risk model, additional flexibility in the age term does not give vastly superior AIC values when compared to the linear age model, although the best in terms of AIC is a model with two internal knots, and hence 4 parameters for the semi-parametric age term. The fitted risks by age for all 27 groups ($3 \times 3 \times 3$ levels of the categorical covariates) for the best model with each link are shown in Figure 5. These results show that the identity link model provides a better fit to the data, suggesting that risk differences are a more appropriate measure than relative risks for this dataset.

Although a monotonic dependence on age is natural for mortality, in this particular analysis it has little effect on the fitted model. Figure 6 displays the fitted age-specific regression functions for both models, showing that the fitted function is virtually identical for the identity link model with and without monotonicity, and is identical for the log link model.

Also shown in Figure 6 is a comparison of two approaches for confidence interval estimation. The first approach, shown by the shaded regions, uses asymptotic normality and the information matrix evaluated at the MLE. The information matrix is obtained as the expected second derivative matrix corresponding to the

Figure 5: Fitted risk of 30-day mortality by age in the ASSENT-2 study for different combinations of MI severity (thin: I, medium: II, thick: III/IV), treatment delay (solid: <2 hours, dashed: 2–4 hours, dotted: >4 hours) and region (black: Western, red: Latin America, blue: Eastern Europe), for (A) identity link and (B) log link models.

Figure 6: (A) Adjusted risk difference and (B) adjusted relative risk associated with age (versus 40 years), with pointwise 95% confidence intervals, estimated using the information matrix (shaded) and bootstrap resampling (dashed lines). The dotted red line in panel (A) shows the estimated risk difference when a monotonicity constraint is applied.

binomial log-likelihood function with probabilities specified by eqs (1) and (2), with either the identity or log link function. The second approach, shown by the dashed lines, uses bootstrapping with pointwise confidence intervals determined using the percentile method. We used 1,000 bootstrap resamples with replacement, and due to the stability of our fitting methods, convergence to the MLE was achieved in every resample. It can be seen that the two methods of confidence interval estimation show very close agreement, which provides some level of support for their use in this analysis.

Adjusted rate differences can also be estimated for this data by using our method for an identity link Poisson model, with semi-parametric adjustment for age. Because in this case all of the patients were

observed for the same period of time, the parameter estimates are very similar to those from the identity link binomial model; however, they have a different interpretation – they are the absolute change in the rate of death per patient-month. A comparison of the parameter estimates from the two models is shown in Table 1.

Table 1: Parameter estimates from identity link binomial (risk difference) and identity link Poisson (rate difference) models on the ASSENT-2 data.

Covariate	Risk difference	Rate difference
Severity (vs. I)		
II	−0.0607	−0.0605
III/IV	−0.2676	−0.2692
Treatment delay (vs. < 2 hours)		
2–4 hours	−0.0024	−0.0025
> 4 hours	−0.0014	−0.0010
Region (vs. Western)		
Latin America	−0.0047	−0.0050
Eastern Europe	−0.0373	−0.0359
Age (vs. 40 years)		
50 years	−0.0045	−0.0045
60 years	−0.0179	−0.0178
70 years	−0.0553	−0.0553
80 years	−0.1534	−0.1541

6 Other methods

The methods described in this paper are fully implemented in R packages `addreg` [27] and `logbin` [28], available from the Comprehensive R Archive Network (CRAN). There are three other notable R packages that provide methods for fitting GAMs: `gam`, `gamlss` and `mgcv`. Each of these allows the models with non-canonical link functions discussed in this paper; however, all employ iterative algorithms involving variants of Fisher scoring or Newton–Raphson, making them subject to instability.

The `gam` function in the `gam` package in R [29] fits GAMs using cubic smoothing splines by employing a local scoring algorithm. This consists of a backfitting (Gauss–Seidel) algorithm for fitting the non-parametric parts of the model within a Newton–Raphson step for updating the parametric parts [30]. The inner loop can be shown to always converge, but the outer loop is only guaranteed to do so if some form of step-size optimization is performed ([7], p. 151). However, the implementation in R does not include any option for step-size modification, and additionally there is no check for the validity of the fitted means. This means that convergence is not guaranteed, and when the method does converge, it may be to a value outside the parameter space. For the 1,000 bootstrap samples used to produce the confidence intervals in Figure 6, the algorithm failed to converge to a valid solution in nearly half (49.1%) of the samples for the identity link model. For the log link model, the algorithm converged in all 1,000 samples, but some of the fitted risks exceeded 1 in every case.

The `gamlss` package [31] provides a method for non-parametric modelling of various parameters of the distribution, including the mean. Similarly to `gam`, its fitting algorithm uses backfitting iterations for the non-parametric parts within Newton–Raphson steps that update the parameter estimates. Unlike `gam`, the user is able to specify the step length for updating parameter estimates, but the function terminates with an error if the update produces invalid fitted values. For the relative risk model fitted on the bootstrap samples from Section 5, `gamlss` converged in only 57 samples when the default step size was used. Convergence in all 1,000 samples was achieved when the step size was made sufficiently small. For the risk difference model, however, `gamlss` did not converge in any of the bootstrap samples regardless of the step size chosen, due to the error caused by violation of the parameter space constraint.

The `mgcv` package [32] provides a flexible `gam` function, allowing for a wide variety of smoothers, as well as methods for automatic selection of the level of smoothing. The fitting method uses a penalized iteratively reweighted least squares algorithm [32], which is not guaranteed to converge, but step-halving is invoked if the penalized deviance increases markedly between iterations, or the estimates move outside the parameter space. Additionally, the user can specify a ridge regression penalty to assist with convergence issues caused by unidentifiable estimates. In order to directly compare the performance of `mgcv`'s `gam` function to our method, we fitted identical unpenalized B-spline models to the ASSENT-2 bootstrap data. We found that `gam` achieved stable convergence whenever the MLE was in the interior of the parameter space, but convergence problems were possible when the MLE was on the parameter space boundary, particularly for the identity link model. The nature of these convergence problems was dependent on the version of `mgcv` that was used. In particular, when using version 1.7 we found that convergence could occur to a sub-optimal boundary point, while in version 1.8 we found that the algorithm could fail to declare convergence when the estimates reached the MLE. This behaviour occurred in a large proportion of our bootstrap replications and persisted even when a stricter convergence criterion and a greater number of iterations were used.

None of these methods support monotonicity constraints on the smooth curves. The `GMBBoost` [13] and `GMonBoost` [14] methods employ likelihood-based boosting techniques to fit GAMs with monotonicity constraints, but in the current implementation only canonical link models are allowed, so we cannot compare them with our approach.

Overall this discussion illustrates that although there are other approaches that could potentially be used for semi-parametric modelling of rate differences, risk differences and relative risks, numerical instability is often an issue. Furthermore, monotonicity constraints for non-canonical models are not available in existing software. The stability and flexibility of our method therefore means that it is a useful addition to existing GAM methodology.

7 Discussion

We have presented a method for smooth semi-parametric adjustment of rate differences, relative risks and risk differences. In general, this can be achieved by using GAMs; however, these effect measures require non-standard link functions and the usual fitting algorithms can fail to converge to the MLE. Our method avoids this by employing variants of existing stable CEM algorithms for fully parametric versions of these models, using B-spline basis functions for the smooth components.

The method is itself a CEM algorithm, and relies on the fact that the EM algorithm will always converge to the MLE within a constrained parameter space. Each constrained parameter space is defined by placing a restriction on the shape of the smooth curve, and by applying the algorithm for each constrained parameter space we are guaranteed to find the overall MLE.

We applied our method to data from the ASSENT-2 clinical trial, showing that semi-parametric adjustment for age provided a better fit than entering age as a linear term in both risk difference and relative risk models. Furthermore, the stability of our fitting algorithm allowed us to use bootstrap resampling to estimate confidence intervals for the semi-parametric relationship. Adjusted rate differences can also be estimated using an identity link Poisson model.

The calculations required at each iteration of the EM algorithms for each method presented here are very simple, although the EM algorithm may take a large number of iterations to converge. The overall computational time required by these methods depends on the number of parameters that must be estimated in a particular model. The models presented in Figure 5 required approximately 3 and 2 minutes, respectively, to find the MLE on a 3.4 GHz processor. One potential method for reducing the computational time is to exploit the fact that the EM algorithms for each parameterization are independent, and could be conducted in parallel on a multi-core processor. Other techniques for speeding up convergence of CEM algorithms have been discussed by Marschner [6].

Further adaptations of our approach are possible. For example, Marschner, Gillett and O'Connell [33] have presented an extension of the CEM algorithm for the identity link Poisson model, in which additional categorical stratification factors have a multiplicative effect on the Poisson means. The algorithm is similar to that for the additive model, but each constrained MLE is found by using an expectation–conditional maximization (ECM) algorithm. It is straightforward to impose non-negativity constraints on the additive parameters, so we can extend this method to allow smooth semi-parametric components by using the same approach described in Section 4.2. An alternative approach to incorporating both additive and multiplicative effects in the same model was provided by the LEXPIT model of Kovalchik, Varadhan, Fetterman, Poitras, Wacholder and Katki [34]. We anticipate that our approach may be useful in extending this model semi-parametrically, although we have not yet investigated this.

Most applications of GAMs use penalized likelihood to allow for flexibility while lessening the tendency to overfit. However, any reasonable penalty term will cause the M-step of the EM algorithm to lose its parameter separation and become a multi-dimensional maximization problem. Marschner and Gillett [5] proposed a solution to this for log binomial models by employing the one-step-late algorithm of Green [35]. Here, the M-step is modified such that parameters associated with the penalty term are replaced by their current estimates. However, this is no longer an EM algorithm, and does not guarantee that the parameter estimates will remain in the parameter space, or even that the likelihood will increase at each step. This can be remedied by a process similar to step-halving, whereby if the new estimate has lower likelihood or is outside the parameter space, we replace it with an estimate between the unpenalized and penalized updates, moving closer to the unpenalized estimate until the conditions are met. Nonetheless, while a penalized likelihood version of our method would be possible, the simple spline-based model used here is likely to provide sufficient flexibility in practice.

Rate differences, relative risks and risk differences are useful in biostatistical settings to provide effect size measures in randomized trials and epidemiological studies. However, the GLMs and GAMs used to estimate these effects are also used in other areas. For example, the identity link Poisson model has been recently used in an ecological study [36], the log link binomial model in a study of socioeconomic status [37] and the identity link binomial model in a study of international politics [38]. This suggests that our method may also have wide applicability outside biostatistics.

Acknowledgements: The authors thank the ASSENT-2 investigators for providing data to the National Health and Medical Research Council Clinical Trials Centre, and the anonymous reviewers for their constructive comments to help improve this paper.

Funding: This research was supported by the Australian Research Council DP110101254.

References

1. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman & Hall, 1989.
2. Marschner IC. glm2: fitting generalized linear models with convergence problems. R J 2011;3:12–15.
3. Marschner IC. Stable computation of maximum likelihood estimates in identity link Poisson regression. J Comput Graph Stat 2010;19:666–83.
4. Donoghoe MW, Marschner IC. Stable computational methods for additive binomial models with application to adjusted risk differences. Comput Stat Data Anal 2014;80:184–96.
5. Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. Biostatistics 2012;13:179–92.
6. Marschner IC. Combinatorial EM algorithms. Stat Comput 2014;24:921–40.
7. Hastie T, Tibshirani R. Generalized additive models, 1st ed. London: Chapman & Hall, 1990.
8. Wood SN. Generalized additive models: an introduction with R. Boca Raton, FL: Chapman & Hall, 2006.
9. SAS Institute Inc. SAS/STAT Software, Version 9.4, Cary, NC, 2013. Available at: <http://www.sas.com/>
10. de Boor C. A practical guide to splines (Applied mathematical sciences). New York: Springer-Verlag, 1978.

11. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996;11:89–121.
12. Ramsay JO. Monotone regression splines in action. *Stat Sci* 1988;3:425–41.
13. Leitenstorfer F, Tutz G. Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* 2007;8:654–73.
14. Tutz G, Leitenstorfer F. Generalized smooth monotonic regression in additive modeling. *J Comput Graph Stat* 2007;16:165–88.
15. Ruppert D, Wand MP, Carroll RJ. Semiparametric regression (Cambridge series in statistical and probabilistic mathematics). New York: Cambridge University Press, 2003:131.
16. Friedman JH, Silverman BW. Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* 1989;31:3–39.
17. Ruppert D. Selecting the number of knots for penalized splines. *J Comput Graph Stat* 2002;11:735–57.
18. Yao F, Lee TCM. On knot placement for penalized spline regression. *J Korean Stat Soc* 2008;37:259–67.
19. Green P, Silverman B. Nonparametric regression and generalized linear models: a roughness penalty approach (Monographs on statistics and applied probability), 1st ed. London: Chapman & Hall, 1994.
20. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974;19:716–23.
21. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. New York: Springer-Verlag, 2002.
22. Hurvich CM, Simonoff JS, Tsai C-L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Stat Soc Ser B* 1998;60:271–93.
23. Malloy EJ, Spiegelman D, Eisen EA. Comparing measures of model selection for penalized splines in Cox models. *Comput Stat Data Anal* 2009;53:2605–16.
24. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977;39:1–38.
25. Baker SG. The multinomial–Poisson transformation. *Statistician* 1994;43:495–504.
26. ASSENT-2 Investigators. Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *Lancet* 1999;354:716–22.
27. Donoghoe M. addreg: additive regression for discrete data. 2014a. Available at: <http://CRAN.R-project.org/package=addreg>, R package version 1.2
28. Donoghoe M. logbin: relative risk regression using the log binomial model. 2014b. Available at: <http://CRAN.R-project.org/package=logbin>, R package version 1.0
29. Hastie TJ. gam: generalized additive models. 2013. Available at: <http://CRAN.R-project.org/package=gam>, R package version 1.09
30. Hastie TJ. Generalized additive models (Chap. 7). In: Chambers JM and Hastie TJ, editors. *Statistical models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1992;249–307.
31. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape (with discussion). *Appl Stat* 2005;54:507–54.
32. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Ser B* 2011;73:3–36.
33. Marschner IC, Gillett AC, O'Connell RL. Stratified additive Poisson models: computational methods and applications in clinical epidemiology. *Comput Stat Data Anal* 2012;56:1115–30.
34. Kovalchik SA, Varadhan R, Fetterman B, Poitras NE, Wacholder S, Katki HA. A general binomial regression model to estimate standardized risk differences from binary response data. *Stat Med* 2013;32:808–21.
35. Green PJ. On use of the EM algorithm for penalized likelihood estimation. *J R Stat Soc Ser B* 1990;52:443–52.
36. Stjernman M, Green M, Lindström Å, Olsson O, Ottvall R, Smith HG. Habitat-specific bird trends and their effect on the farmland bird index. *Ecol Indic* 2013;24:382–91.
37. Hiyoshi A, Fukuda Y, Shipley MJ, Bartley M, Brunner EJ. A new theory-based social classification in Japan and its validation using historically collected information. *Soc Sci Med* 2013;87:84–92.
38. Berger D, Corvalan A, Easterly W, Satyanath S. Do superpower interventions have short and long term consequences for democracy? *J Comp Econ* 2013;41:22–34.