1 Modelling Strategies and Guidelines for GAITD Regression

Fitting a suitable GAITD regression model to a count response involves many decisions. The following general tips and strategies are suggested. The skill set required for some steps are much higher than others.

1.1 Background

To start, it is assumed that the reader is familiar with the following background material.

(i) The notation and basic GAITD regression formulas, such as the special values

$$\mathcal{S} \;=\; \{\mathcal{A}_p,\ \mathcal{A}_{np},\ \mathcal{I}_p,\ \mathcal{I}_{np},\ \mathcal{T},\ \mathcal{D}_p,\ \mathcal{D}_{np}\} \tag{1}$$

and $\boldsymbol{\eta}^T = (\eta_1, \ldots, \eta_M)^T =$

$$\begin{pmatrix} g_\pi(\theta_\pi),\ \log\frac{\omega_p}{\mathcal{N}},\ g_\alpha(\theta_\alpha),\ \log\frac{\phi_p}{\mathcal{N}},\ g_\iota(\theta_\iota),\ \log\frac{\psi_p}{\mathcal{N}},\ g_\delta(\theta_\delta),\ \log\frac{\omega_1}{\mathcal{N}},\ \ldots,\ \log\frac{\omega_{|\mathcal{A}_{np}|}}{\mathcal{N}}, \\ \log\frac{\phi_1}{\mathcal{N}},\ \ldots,\ \log\frac{\phi_{|\mathcal{I}_{np}|}}{\mathcal{N}},\ \log\frac{\psi_1}{\mathcal{N}},\ \ldots,\ \log\frac{\psi_{|\mathcal{D}_{np}|}}{\mathcal{N}} \end{pmatrix}, \tag{2}$$

where $g_\cdot(\cdot)$ are the link functions applied to $f_\pi,\ f_\alpha,\ f_\iota,\ f_\delta$. The quantity $\mathcal{N} = 1 - \omega_p - \phi_p - \psi_p - \sum\omega_u - \sum\phi_u - \sum\psi_u$ corresponds to the multinomial logit model reference group.

(ii) The 1-parameter combo PMF is $\Pr(Y = y; \theta_\pi, \omega_p, \theta_\alpha, \phi_p, \theta_\iota, \psi_p, \theta_\delta, \boldsymbol{\omega}_{np}, \boldsymbol{\phi}_{np}, \boldsymbol{\psi}_{np}) =$

$$\begin{cases} 0, & y \in \mathcal{T}, \\ \omega_p\, f_\alpha(y) \,/ \sum\limits_{u\in\mathcal{A}_p} f_\alpha(u), & y \in \mathcal{A}_p, \\ \omega_s, & y = a_s \in \mathcal{A}_{np},\ s = 1, \ldots, |\mathcal{A}_{np}|, \\ \Delta\, f_\pi(y) + \phi_p\, f_\iota(y) \,/ \sum\limits_{u\in\mathcal{I}_p} f_\iota(u), & y \in \mathcal{I}_p, \\ \Delta\, f_\pi(y) + \phi_s, & y = i_s \in \mathcal{I}_{np},\ s = 1, \ldots, |\mathcal{I}_{np}|, \\ \Delta\, f_\pi(y) - \psi_p\, f_\delta(y) \,/ \sum\limits_{u\in\mathcal{D}_p} f_\delta(u), & y \in \mathcal{D}_p, \\ \Delta\, f_\pi(y) - \psi_s, & y = d_s \in \mathcal{D}_{np},\ s = 1, \ldots, |\mathcal{D}_{np}|, \\ \Delta\, f_\pi(y), & y \in \mathcal{R}\backslash\mathcal{S}, \end{cases} \tag{3}$$

where Δ is a normalizing constant.

(iii) VGLMs and its infrastructure, especially $\boldsymbol{\eta}$, constraint matrices $\mathbf{H}_k$, and the multinomial logit model (MLM) with its softmax function. VGAMs are needed for smoothing. Yee and Ma (2024) is probably the best for these topics. For convenience, the following formulas are a brief summary of most of these.

For most VGLMs the data can be written $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, independently and the PMF/PDF for the ith observation is

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) \;=\; h(\boldsymbol{y}_i; \eta_1(\boldsymbol{x}_i), \ldots, \eta_M(\boldsymbol{x}_i))$$

for some $f(\cdot)$ and $h(\cdot)$. The $\boldsymbol{\theta}$ are generic parameters, and coupled with $\eta_j = g_j(\theta_j) = \boldsymbol{\beta}^T\boldsymbol{x}$, the parameters are modelled as linear predictors. The parameter link functions g_j are used to

transform the parameters. One can write $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ or $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T$. For most VGLMs the log-likelihood $\ell = \sum_{i=1}^n w_i^* \, \ell_i(\eta_1, \ldots, \eta_M)$ is maximized. The w_i^* are known positive prior weights.

For VGLMs:

$$\boldsymbol{\eta}(\boldsymbol{x}) \;=\; \mathbf{H}_1 \, \boldsymbol{\beta}_{(1)}^* x_1 + \cdots + \mathbf{H}_d \, \boldsymbol{\beta}_{(d)}^* x_d \;=\; \mathbf{B}^T \boldsymbol{x} \tag{4}$$

where $\mathbf{H}_1, \ldots, \mathbf{H}_d$ are *known* full-column rank *constraint matrices*, and $\boldsymbol{\beta}_{(k)}^*$ is a vector containing a possibly reduced set of unknown regression coefficients. With no constraints at all, $\mathbf{H}_k = \mathbf{I}_M$ for all k. Usually $x_1 = 1$ (intercept term). In general,

$$\mathbf{B}^T \;=\; \left(\mathbf{H}_1 \boldsymbol{\beta}_{(1)}^* \;\; \cdots \;\; \mathbf{H}_d \boldsymbol{\beta}_{(d)}^* \right). \tag{5}$$

For VGAMs: (4) extends to

$$\boldsymbol{\eta}(\boldsymbol{x}) \;=\; \mathbf{H}_1 \, \boldsymbol{\beta}_{(1)}^* x_1 + \mathbf{H}_2 \, \boldsymbol{f}_2^*(x_2) + \cdots + \mathbf{H}_p \, \boldsymbol{f}_d^*(x_d) \tag{6}$$

where $\boldsymbol{f}_k^*(x_k) = (f_{(1)k}^*(x_k), \ldots, f_{(r_k)k}^*(x_k))^T$ is a r_k-vector of smooth functions of x_k (estimated by a vector smoothing spline). With no constraints, $\eta_j = \sum_{k=1}^{p} f_{(j)k}(x_k)$.

For RR-VGLMs:

$$\boldsymbol{\eta}(\boldsymbol{x}) \;=\; \mathbf{B}_1^T \boldsymbol{x}_1 + \mathbf{A} \, \boldsymbol{\nu} \tag{7}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$, $\boldsymbol{\nu} = \mathbf{C}^T \boldsymbol{x}_2$ is an R-vector of latent variables, $\mathbf{A}$ is $M \times R$ and $\mathbf{C}$ is $p_2 \times R$. Note that $\mathbf{A}$ and $\mathbf{C}$ are general (thin) matrices whereas *doubly constrained* RR-VGLMs (DRR-VGLMs) allow the matrices to have structure.

For DRR-VGLMs:

$$\boldsymbol{\eta} \;=\; \mathbf{B}_1^T \boldsymbol{x}_1 + \left\{ \sum_{r=1}^{R} \boldsymbol{e}_r^T \otimes \begin{pmatrix} \boldsymbol{e}_r \\ \widetilde{\mathbf{H}}_{\mathrm{A}r} \, \widetilde{\boldsymbol{a}}_r^* \end{pmatrix} \right\} \left\{ \sum_{k=1}^{p_2} \boldsymbol{e}_k^T \otimes (\mathbf{H}_{\mathrm{C}k} \, \boldsymbol{c}_k^*) \right\} \boldsymbol{x}_2. \tag{8}$$

Here, $\dim(\boldsymbol{x}) = d = p$ with $\dim(\boldsymbol{x}_1) = p_1$, $\dim(\boldsymbol{x}_2) = p_2$, and $p_1 + p_2 = d$. Also, $\mathbf{A}$ and $\mathbf{C}$ are estimated, and $\mathbf{B} = (\mathbf{B}_1^T \;\; \mathbf{B}_2^T)^T$ with $\mathbf{B}_2 = \mathbf{C} \, \mathbf{A}^T$, a reduced-rank approximation of a subset of $\mathbf{B}$ (cf. (4)). The *rank* R is often 1 or 2, maybe 3 .... Using corner constraints is one method to ensure $\mathbf{A}$ and $\mathbf{C}$ are unique.

For the MLM:

$$g(p_s) \;=\; \eta_s \;=\; \log \{p_s / p_{D+1}\}, \qquad s = 1, \ldots, D,$$

where $\boldsymbol{p} = (p_1, \ldots, p_D)^T$ is a vector of probabilities. Then g is known as the multilogit link, and $p_{D+1} = 1 - \sum_{u=1}^{D} p_u$ corresponds to the reference/baseline group. The inverse link (softmax) is $p_s = e^{\eta_s} / \sum_{u=1}^{D+1} e^{\eta_u}$ where $\eta_{D+1} \equiv 0$ for identifiability.

(iv) Prior experience using the VGAM package is ideal, especially `vglm()`, `multinomial()`, the `zero` argument and `negbinomial()`. The help file for `gaitdpoisson()` really needs to be digested. Yee (2015) and Yee (2008) are good for these topics, as is VGAMrefcard.pdf at `https://www.stat.auckland.ac.nz/~yee`

(v) Firstly, the Shiny app at `https://www.stat.auckland.ac.nz/~yee` is strongly recommended for exploring the combo PMF interactively and tying together (3) with the software.

(vi) Some experience fitting some simple special cases, such as the zero-inflated Poisson model, certainly helps but is not necessary.

We split the following items into three stages.

1.2 (I) Pre-fitting

1. Are the data are representative of future data sets? If not, there is little point in fitting a regression model for prediction. If not, a regression model can still be very useful for 'explaining' the data, e.g., determine which variables are associated with the response, identifying certain subgroups, and how much measurement error there is.

2. Are the data heaped and/or seeped? That is, is there measurement error?

3. Partition the data into training and test sets and spikeplot them. Repeating this a few times, they should give an idea how sampling variation affects certain features such as spikes and dips (holes). This helps in the choice of $\mathcal{A}$, $\mathcal{I}$, $\mathcal{D}$ below. One should try to have as few special values assigned to these sets as possible. Ignore features that sometimes vanish, e.g., small spikes.

4. Choose a parent distribution based on the spikeplots. This depends on the general shape and/or quantities such as the variance-to-mean ratio. For a long monotonic tail, the zeta and logarithmic can be suitable. If overdispersed and unimodal then possibly a negative binomial (NB), else a Poisson. If there are no special values exhibiting a nonignorable feature then a standard regression model could suffice; GAITD regression would add little benefit for much complexity.

5. Identify any truncated values for $\mathcal{T}$ first. These come from three regions: the lower tail, upper tail, and special values in between. There ought to be a good explanation for each value, such as a structural reason. Assigning a value for $\mathcal{T}$ just because there are no such values in the data set is a weak and dubious reason.

6. If present, starting with the largest spikes (or heaps), specify $\mathcal{A}$ or $\mathcal{I}$ (or $\mathcal{D}$ later) depending on the research question. Recall that GA can explain why observations are there, GI accounts for why they are there in excess, and GD regression can explain why observations are not there. Ideally there is external justification for each element in $\mathcal{A}$ and $\mathcal{I}$ to strengthen the overall choice and make the model more defensible, e.g., in the case of heaping/seeping.

   Depending on the specific research question(s), choose between $\mathcal{A}$, $\mathcal{I}$ and $\mathcal{D}$. If unsure, choose $\mathcal{A}$.

7. We recommend choosing which from the four operators first based on the research question, and then its values. After that, decide between parametric and nonparametric.

8. If present, do likewise for dips and $\mathcal{D}$ last. Deflation is much harder to model compared to alteration and inflation. From item 3, ignore minor spikes and dips that may be due to sampling variation or are practically unimportant. It only takes one misspecified special value for numerical problems to occur.

9. Outliers on the RHS tail and other features not able to be handled might be *artificially* censored by altering, e.g., Yee and Ma (2024) has an example. But the action and consequences need to be described in the report.

10. Choose between parametric and nonparametric A/I/D—do the spikes/dips follow a similar distribution as the parent? Parametric is preferred but this should be checked by using training and test data. Overfitting should be borne in mind.

11. If $\mathcal{A}_p$, $\mathcal{I}_p$, $\mathcal{D}_p$ are specified then ideally their cardinality is large enough: $|\mathcal{A}_p| > \dim(\boldsymbol{\theta}_\pi)$, $|\mathcal{I}_p| > \dim(\boldsymbol{\theta}_\pi)$, $|\mathcal{D}_p| > \dim(\boldsymbol{\theta}_\pi)$. One borrows strength by having $\boldsymbol{\theta}_\pi = \boldsymbol{\theta}_\alpha = \boldsymbol{\theta}_\iota = \boldsymbol{\theta}_\delta$ (arguments `eq.ap`, `eq.ip`, `eq.dp`) so that estimation is much easier and stable. If not, then the position of the elements in $\mathcal{A}_p$, $\mathcal{I}_p$, $\mathcal{D}_p$ determine the overall stability of the model, e.g., if all the elements in $\mathcal{A}_p$ are close to each other relative to $f_\alpha(\cdot)$ then $\text{Var}(\widehat{\boldsymbol{\theta}}_\alpha)$ will be very large. Hence the range of elements within each of $\mathcal{A}_p$, $\mathcal{I}_p$, $\mathcal{D}_p$ ought to be wide.

    In passing, it is noted that although multicollinearity applies to the columns of $\mathbf{X}$ in GLMs, the problems described here involving values of $\boldsymbol{y}$ bear some similarity.

1.3 (II) Fitting

12. The Shiny app may be helpful for item 10.

13. Fit a null model first. Set `trace = TRUE` because the performance of the algorithm gives important insights into the underlying problem modelling (Osborne, 1992). Well-conditioned VGLMs usually converge within 7–9 IRLS iterations. Any numerical problems are indicative of misspecified $\mathcal{S}$ values. If needed, the Shiny app is useful for inputting initial values.

    Too much A/I/D will lead to the normalizing constant Δ becoming negative. For example, the combined effects of $\mathcal{A}_{np}$, $\mathcal{I}_{np}$, $\mathcal{D}_{np}$ are particularly expensive so that there is no baseline probability left. Sometimes it is necessary to analyze a subset of the data, e.g., $Y > 0$ when the sample proportion of 0s is high.

14. When adding covariates, automatic methods for variable selection are better applied to a few preselected regressors than feeding in a large vector $\boldsymbol{x}$ without thought. Although the MLM is intercept-only by default, care is needed fitting the MLM with covariates—the number of coefficients multiplies quickly so there may be interpretability problems. Reduced rank vector generalized linear models (RR-VGLMs) and doubly constrained RR-VGLMs (DRR-VGLMs) are potentially suitable but require finesse.

    If the intercepts are to be interpreted (a useful feature to have), centering the covariates is a good idea. But `multilogitlink(, inverse = TRUE)` is not easy to compute in one's head.

15. For $\mathcal{A}_{np}$, $\mathcal{I}_{np}$, $\mathcal{D}_{np}$, if some MLM probabilities appear equal then a parallelism constraint might be applied. Arguments `parallel.a`, `parallel.i`, `parallel.d` support this. However, it is difficult to apply a parallelism constraint to a subset of MLM probabilities. One would have to use `constraints` to manually input them, and constructing constraint matrices would entail laborious work and meticulous bookkeeping and care based on (2).

16. With covariates, variable selection with `step4()` applied to a `"vglm"` object is a possibility because stepwise regression based on AIC is more theoretically grounded with parametric models. In contrast, additive models are more heuristic and inference is based on approximations and asymptotics.

17. If there are only a very few regressors, try fitting an additive model, i.e., `vgam()`. Smoothers can suggest transformations or terms easily replaced by low-degree polynomials, etc.

1.4 (III) Post-fitting

18. Rootograms (Kleiber and Zeileis, 2016) are a useful diagnostic tool: see `?rootogram4`.

19. Another diagnostic is `simulate(fit)` and then overlay each simulation on a spikeplot of the original data or a test set.

20. Be aware of the limitations of p-values. Because the choice of $\mathcal{A}$, $\mathcal{I}$, $\mathcal{T}$, $\mathcal{D}$ usually involves looking at the data, any p-values should be cited sparingly (if at all) and interpreted with caution and admission. A caveat is often appropriate. This is an open research area.

21. Use training and test data to validate the above. Guard against overfitting since a slightly underfitting model is preferred over a mildly overfitting one. Because if its immense flexibility, overfitting remains probably the biggest problem amateurs will face with GAITD regression.

Last modified: 2024–05.

1.5 Further instructions

1. Based on a few spikeplots from item 3, identify and sort the major features into a list in descending order. There should be a maximum of seven features, as there are only seven(!) operators. Ideally there are far fewer. Guard against overfitting.

2. When assigning elements into each of the seven subsets comprising $\mathcal{S}$, ask: is there any plausible reason to explain this?

3. Going down the list, assign each feature to all remaining operators. That is,

    - choose between parametric and nonparametric, with the first being preferable and nonparametric handling a few aberrant values that are more inexplicable.
    - Consider changing from one operator to another, as the number of choices diminishes. For example, replace $\mathcal{A}_p$ to $\mathcal{I}_p$. The most important features should be handled by the operator(s) that answers the main research question(s). Nuisance features may have to be handled using an operator that models the effect but doesn't answer any research questions per se.
    - If there are no operators left, then are the remaining features ignorable? If not, try repeating these instructions using a different permutation of choices.