gamdist: Generalized Additive Models in Python

Bob Wilson

August 31, 2018

Abstract

TBD

Contents

1 Introduction

This paper introduces a Python library called `gamdist`, which uses a distributed optimization technique called the Alternating Direction Method of Multipliers (ADMM) to fit a special type of regression model called a Generalized Additve Model (GAM) to data.

Outline of Paper In §2 we describe Generalized Additive Models. In §3 we describe the Alternating Direction Method of Multipliers and how it may be used to fit GAMs. In §4, we describe the architecture of the library, including relevant implementation details.

2 Generalized Additive Models

The primary goal of `gamdist` is the estimation of certain aspects of the joint distribution of a collection of one or more random variables X called *features* and a random variable Y we will call the *response*. Specifically, we are interested in the conditional distribution of $Y \mid X$. Perhaps the simplest approach is the linear model, which assumes $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T\beta$. This notation captures three key assumptions of the linear model. First, the conditional distribution is Gaussian for all values of X. Second, the mean of the distribution depends on the features X in a fairly specific way discussed below. Third, the variance is the same for all values of X. These assumptions are all loosened in various generalizations of the linear model used in `gamdist`.

If we choose $\nu(x) = x$, then the assumption is that $\mu(x) = x^T\beta$, and the mean depends linearly on the features; however, the *linear* in *linear model* refers to the dependence on β, not on the features. It is common to include a constant term in $\nu(x)$ to account for an affine dependency between the features and the response. For example, $\nu(x) = \begin{bmatrix}1 & x_1 & x_2 & \cdots\end{bmatrix}^T$. We might incorporate quadratic terms to capture nonlinear dependencies including interactions, such as $\nu(x) = \begin{bmatrix}x_1 & x_2 & x_1^2 & x_1 \cdot x_2 & x_2^2 & \cdots\end{bmatrix}^T$. Or we might include more exotic transformations of the features, like $\nu(x) = \begin{bmatrix}\log(x_1) & \sin(x_2) & \cdots\end{bmatrix}^T$. These models are nonlinear in the features, but linear in the parameters β; however, the linear model will only incorporate transformations that we explicitly include.

Another common situation is when some or all of the features are categorical. For example, consider a model with a single feature corresponding to a person's favorite color, and suppose choices are limited to red, green, and blue. The model would consist of the average responses for people who prefer any particular color. We might support such a model by defining

$$\nu(x) = \begin{cases} \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T & \text{if } x = \text{red} \\ \begin{bmatrix}0 & 1 & 0\end{bmatrix}^T & \text{if } x = \text{green} \\ \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T & \text{if } x = \text{blue.} \end{cases}$$

Alternative approaches to encoding categorical variables are common and useful in different

circumstances. We see that the simple linear model is applicable to a wide range of problems, even those that may not appear linear at first glance.

Fitting a linear model to a set of observations is called linear regression, and is accomplished by solving a least squares optimization problem. This is an example of maximum likelihood estimation (MLE), itself a special case of maximum a posteriori (MAP) estimation, which is the unifying approach used throughout `gamdist`. It is worth formulating this optimization problem so that we may see how it evolves as we consider more general scenarios.

Suppose we have n observations of the form $\{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$. We assume these observations are independent and drawn from the same (unknown) joint distribution.[1] Under the assumptions of the linear model, $Y \mid X = x^{(i)} \sim \mathcal{N}(\mu(x^{(i)}), \sigma^2)$. The likelihood of a particular observation is

$$\mathcal{L}(\beta; x^{(i)}, y^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2\right).$$

Note that by convention the likelihood is interpreted as a function of β parameterized by the observation $(x^{(i)}, y^{(i)})$. The likelihood of the entire set of observations is the product of the likelihoods of the individual observations: $\mathcal{L}(\beta; x, y) = \Pi_{i=1}^{n} \mathcal{L}(\beta; x^{(i)}, y^{(i)})$. The log-likelihood is the sum of the log-likelihoods of the individual observations:

$$\ell(\beta; x, y) = \log \mathcal{L}(\beta; x, y) = -n/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2.$$

Maximizing the likelihood is the same as maximizing the log-likelihood, and if we are only interested in estimating β, this is equivalent to the problem

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2,$$

where the variable is β and $y^{(i)}$ and $\nu(x^{(i)})$ are data. An elementary result in optimization theory is that a unique solution exists if and only if $V^T V$ is full rank, where the ith row of V is equal to $\nu(x^{(i)})^T$. In that case, the optimal β satisfies the so-called normal equations:

$$V^T V \hat{\beta} = V^T y.$$

If we assume the model is correct (that is, that the conditional distribution really has the assumed form), exact formulae exist for confidence intervals on the parameters β. If the variance is unknown, it too can be estimated from the data. We can employ hypothesis tests against the null hypothesis that some or all of the components of β are zero. We can apply the resulting model to new data assumed to be drawn from the same joint distribution to compute confidence intervals on the response, $Y \mid X = x_{\text{new}}$ or the mean of this distribution,

[1] This assumption is loosened in random effects or mixed effects models which are not considered here; see [Str12].

$\mu(x_{\text{new}})$. These are immensely valuable tools in the analysis of data and the application of data to predictions. A good reference on linear models is [Wei05]. Useful results are collected in Appendix A.

We can loosen the assumption of constant variance slightly without materially affecting the inferences that may be drawn. Suppose the ith observation has variance $\sigma^2/\phi^{(i)}$, where $\phi^{(i)}$ is a known quantity, but σ^2 may or may not be known a priori.

Even if the assumptions underlying the linear model are correct, if the noise is high, or if the number of features is large relative to the number of observations, we may use regularization to improve both the estimates of β and predictions based on the estimated model. Regularization reduces the sensitivity of the estimates to noise at the expense of introducing bias, and may be thought of as imposing a Bayesian prior on the parameters β. Some of the most common forms of regularization include ridge regression and the lasso [Tib96]. For example, the lasso may be formulated as the problem:

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu \left(x^{(i)} \right)^T \beta \right)^2 + \lambda \cdot \|\beta\|_1,$$

but this problem does not have a closed-form solution. Moreover, introducing regularization means the formulation is no longer a maximum likelihood estimation problem. Instead, it is a maximum a priori estimation problem. MLE problems have some statistical properties that MAP estimation problems do not possess.

If the constant variance assumption does not hold, but the dependence is known, then Weighted Least Squares.

If the conditional distribution is not Gaussian, other techniques may prove more useful. For example, if the conditional distribution is Laplacian, we may use least absolute deviation regression instead of least squares [BD93]. Like with the lasso, there are no exact formulae for statistical inference in this context and we must settle for an asymptotic or non-parametric approach such as the bootstrap [ET93].

Yet another approach to extending linear models was introduced by [NW72] and discussed in detail in [MN89]. Their formulation extends the linear model in a few ways. The conditional distribution is not assumed to be Gaussian. Common alternatives include the binomial and Poisson distributions. The mean of the distribution is permitted to depend on the features in a more complicated way, via the introduction of a *link function*, g: $\mu(x) = g^{-1}(\eta(x))$, where $\eta(x) = \nu(x)^T\beta$. When $g(x) = x$, this recovers the same relationship between the features and μ assumed in the linear model, but other link functions may be used like the logistic function $g(x) = \log(x/(1-x))$. Finally, the variance is sometimes permitted to depend on x instead of being constant. Such models are called Generalized Linear Models (GLMs). Fitting such models is accomplished via maximum likelihood estimation:

$$\text{minimize} \quad \sum_{i=1}^{n} \ell \left(\beta; x^{(i)}, y^{(i)} \right) + r(\beta),$$

where $r(\beta)$ is a regularization term on β, such as in the lasso. For many choices of distribution family and link function, the corresponding optimization problem is convex.

Introduced by [HT86], Generalized Additive Models (GAMs) extend GLMs by permitting $\eta(x)$ to be a nonparametric function of the features: $\eta(x) = \sum_{i=1}^{p} h_i(x_i)$, where h_i are smooth functions. When $h_i(x_i) = \beta_i x_i$, the linear model is recovered (all of the parametric dependencies discussed with regards to ν are still possible here of course, but the idea is that the data itself should tell us the form of the relationship). Typically the h_i functions are chosen to be some sort of spline, such as a natural cubic spline. GAMs are also fit via MLE, and many practical problems can be formulated as convex optimization problems.

(Reference GAM, GAMr, Casella and Berger, Seber) regularized GAMs

3 The Alternating Direction Method of Multipliers

4 Software Architecture

Acknowledgments

This is an example of an unnumbered section.

A Properties of the Linear Model

In this section, we state various useful facts (without proof) regarding the linear model. For details, see [Wei05], [SL03], [Woo17], and [CB01]. The goal of this section is twofold: to gather formulae used in `gamdist`, and to illustrate that regression analysis does not end with model fitting. In fact, it is often desirable to quantify the uncertainty in fitted model parameters, check the assumptions of the linear model through examination of the residuals, perform model selection, and quantify the uncertainty associated with predictions. These are worthwhile goals in any regression analysis.

Suppose $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T\beta$ and $\nu(x) \in \mathbf{R}^p$ is a known function. Let $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ be IID samples drawn from the joint distribution of X and Y. Let V be the matrix whose ith row is $\nu(x^{(i)})$, and assume V is full rank. Let $\hat{\beta} = (V^TV)^{-1}V^Ty$. Then

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(V^TV)^{-1}). \tag{1}$$

Specifically, $\hat{\beta}$ has a multivariate normal distribution, and $\hat{\beta}$ is an unbiased estimate of β. It is also a *consistent* estimate of β, meaning that $\hat{\beta}$ converges in probability to β, as n increases without limit. It is also the best linear unbiased estimate of β: any other linear, unbiased estimates have higher variance than $\hat{\beta}$ [Woo17, § 1.3.9].

If σ^2 is unknown, it may be estimated from the data. Let

$$\hat{\sigma}^2 = \frac{\|y - V\hat{\beta}\|_2^2}{n - p}. \tag{2}$$

Then $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ [Wei05, § 3.4.4]. This indicates that $\hat{\sigma}^2$ is an unbiased estimate of σ^2. Suppose $\mathbf{Prob}\{\chi^2_{n-p} \in (\ell, u)\} = \alpha$; that is, (ℓ, u) is a confidence interval (at level α) on a χ^2_{n-p} random variable. Then

$$\mathbf{Prob}\left\{\sigma^2 \in \left(\frac{(n-p)\cdot\hat{\sigma}^2}{u}, \frac{(n-p)\cdot\hat{\sigma}^2}{\ell}\right)\right\} = \alpha. \tag{3}$$

A particularly useful case is when $u \to \infty$, corresponding to $\ell = \Phi^{-1}_{\chi^2_{n-p}}(1-\alpha)$, where $\Phi_{\chi^2_{n-p}}$ is the cumulative distribution function of a χ^2_{n-p} random variable, in which case

$$\mathbf{Prob}\left\{\sigma^2 < \frac{(n-p)\cdot\hat{\sigma}^2}{\Phi^{-1}_{\chi^2_{n-p}}(1-\alpha)}\right\} = \alpha,$$

corresponding to an upper confidence limit on σ^2, at level α.

We may compute confidence intervals on linear combinations of the components of β by noting that $c^T\hat{\beta} \sim \mathcal{N}(c^T\beta, \sigma^2 \cdot c^T(V^TV)^{-1}c)$, and thus

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\sigma^2 \cdot c^T(V^TV)^{-1}c}} \sim \mathcal{N}(0,1).$$

Let $\pm z_\alpha$ be the endpoints of a confidence interval on a standard Gaussian random variable at level α. Then

$$c^T\hat{\beta} \pm z_\alpha \cdot \sqrt{\sigma^2 \cdot c^T(V^TV)^{-1}c} \tag{4}$$

are the endpoints of a confidence interval on $c^T\beta$. This formula is only computable when σ^2 is known a priori. When σ^2 is unknown, we need a formula in terms of its estimated value, $\hat{\sigma}^2$, leading to

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\hat{\sigma}^2 \cdot c^T(V^TV)^{-1}c}} \sim t_{n-p}.$$

Thus, when σ^2 is unknown, a confidence interval on $c^T\beta$, at level α, has endpoints

$$c^T\hat{\beta} \pm t_{n-p;\alpha} \cdot \sqrt{\hat{\sigma}^2 \cdot c^T(V^TV)^{-1}c}, \tag{5}$$

where $\pm t_{n-p;\alpha}$ are the endpoints of a confidence interval, at level α, on a Student's t distribution with $n-p$ degrees of freedom.

These facts can be used for testing $H_0 : c^T\beta = 0$ vs. the alternative, $H_1 : c^T\beta \neq 0$ for a particular value of c. Under the null hypothesis,

$$T_c := \frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 \cdot c^T(V^TV)^{-1}c}} \sim t_{n-p}, \tag{6}$$

so the p-value associated with the test is simply $\Phi_{t_{n-p}}(-|T_c|) + (1 - \Phi_{t_{n-p}}(|T_c|))$, where $\Phi_{t_{n-p}}$ is the cumulative distribution function for a Student's t distribution with $n-p$ degrees of freedom.

As special cases of the above discussion, when $c = \hat{e}_i$, we get confidence intervals for, and hypothesis tests regarding, β_i. When $c = \nu(x_{\text{new}})$, we get confidence intervals for $\mu(x_{\text{new}})$; that is, the mean response of the model applied to a new data point. Confidence intervals on $Y \mid X = x_{\text{new}}$ involve an extra component of uncertainty due to the variance of the conditional distribution: even if we knew $\mu(x_{\text{new}})$ perfectly, the conditional distribution still has variance σ^2. This leads to an extra term of $\hat{\sigma}^2$ in Equation (5):

$$\nu(x_{\text{new}})^T \hat{\beta} \pm t_{n-p;\alpha} \cdot \sqrt{\hat{\sigma}^2 \cdot (1 + \nu(x_{\text{new}})^T (V^T V)^{-1} \nu(x_{\text{new}}))},$$

which are the endpoints of a confidence interval on Y [Wei05, §3.6].

When we want simultaneous confidence intervals on multiple linear combinations of β, we must adjust for the multiple comparisons. There is more than one approach to doing so. Suppose we are interested in $C\beta$, where $C \in \mathbf{R}^{q \times p}$. Then

$$C\hat{\beta} \sim \mathcal{N}\left(C\beta, \sigma^2 \cdot C(V^T V)^{-1} C^T\right),$$

which shows that $C\hat{\beta}$ is an unbiased estimator for $C\beta$, and that $C\hat{\beta}$ is normally distributed. Note that when $C = V$, we get the distribution of the fitted means, $\hat{\mu}$, since $\hat{\mu} := V\hat{\beta}$. Suppose we are interested in simultaneous confidence intervals on $C\beta$ at level α. The Bonferroni correction defines $\alpha' = 1 - (1 - \alpha)/q$ and then simply uses the endpoints in (5) for each individual component of $C\beta$, substituting α' for α. For example, suppose we wanted a simultaneous 95% confidence interval on $C\beta$, where C has $q = 5$ rows. Then $\alpha = 0.95$ and $\alpha' = 0.99$. So we would compute 99% confidence intervals on each component $c^{(i)}\beta$, where $c^{(i)}$ is the ith row of C. This approach is simple but overly conservative in many cases [Wei05, § 9.1.3].

Tukey's method; Scheffe's method.

Now suppose C is full rank, and that $q < p$. We would like to test the hypothesis $C\beta = d$. As derived in [Woo17, § 1.3.4],

$$T_C := \frac{1}{q}(C\hat{\beta} - d)^T (\hat{\sigma}^2 \cdot C(V^T V)^{-1} C^T)^{-1} (C\hat{\beta} - d) \sim F_{q,n-p},$$

where $F_{q,n-p}$ is Snedecor's F distribution with q and $n-p$ degrees of freedom in the numerator and denominator, resp. This relationship generalizes (6) since an F distribution with one degree of freedom in the numerator is equivalent to a t^2 distribution [Wei05, § 3.5.3]. The p-value for this test would be $\Phi_{F_{q,n-p}}(-|T_C|) + (1 - \Phi_{F_{q,n-p}}(|T_C|))$, where $\Phi_{F_{q,n-p}}$ is the cumulative distribution function for an F distribution with the specified degrees of freedom.

This test statistic is useful in several situations. Suppose we are considering a sequence of nested models: a model consisting only of a grand mean (in which μ does not depend on the features at all), a model consisting only of main effects, a model with first-order interactions, and a model with higher-order interactions. We may wish to check H_0: the model consists only of main effects vs. H_1: interactions are present. This amounts to checking whether $C\beta = 0$, where $C\beta$ are the components of β corresponding to the interaction terms. Or we may want to check $H_0 : \mu(x) = \mu$ vs. $H_1 : \mu(x) \neq \mu$, where μ is the grand mean. In

this case, we are checking whether there is any evidence of the mean response depending on the features. Of course, checking multiple hypotheses is subject to the multiple comparisons problem discussed above.

If a particular feature is categorical with at least three levels, it will consist of at least two parameters. We would typically want to test whether all associated parameters are non-zero, not just one. Or if $\nu(x)$ consists of multiple transformations of a particular feature, like $\nu(x) = \begin{bmatrix} \cdots & x_1 & x_1^2 & \log(x_1) & \cdots \end{bmatrix}^T$, we might want to simultaneously test all the corresponding components of β for being non-zero. The F-test described here is applicable to these scenarios.

What about simultaneous confidence intervals on multiple components of β? How do we do power calculations? What about the residuals, $y^{(i)} - \hat{\mu}^{(i)}$? What do we do with the residuals (graph them against the mean response, make a Q-Q plot, standardize them, etc.)

References

[BD93] D. Birkes and Y. Dodge. *Alternative Methods of Regression.* John Wiley & Sons, Inc., 1993.

[CB01] G. Casella and R. Berger. *Statistical Inference.* Cengage Learning, 2nd edition, 2001.

[ET93] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, 1993.

[HT86] T. Hastie and R. Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 1986.

[MN89] P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman & Hall, 2nd edition, 1989.

[NW72] J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

[SL03] G. Seber and A. Lee. *Linear Regression Analysis.* John Wiley & Sons, Inc., 2nd edition, 2003.

[Str12] W. Stroup. *Generalized Linear Mixed Models.* Chapman & Hall, 2012.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[Wei05] S. Weisberg. *Applied Linear Regression.* John Wiley & Sons, Inc., 3rd edition, 2005.

[Woo17] S. Wood. *Generalized Additive Models, An Introduction with R.* Chapman & Hall, 2nd edition, 2017.