Penalized likelihood

Patrick Breheny

December 3, 2025

Introduction

- Today we discuss a very flexible extension known as penalized likelihood
- The basic idea of penalized likelihood is that instead of maximizing the likelihood itself, we will maximize

$$q(\boldsymbol{\theta}|X) = \ell(\boldsymbol{\theta}|X) - p(\boldsymbol{\theta}),$$

  where the penalty function p penalizes what we would think of as unreasonable or unrealistic values of $\boldsymbol{\theta}$ (note that the penalty function does not depend on the data)
- *Note: penalized likelihood is usually presented in terms minimizing the negative log-likelihood ("loss function") as opposed to maximizing the likelihood*

What to penalize?

- What do we mean by unreasonable values of $\boldsymbol{\theta}$?
- We will see a variety of examples in today's lecture, but typically we mean "extreme" values, such as infinite regression parameters, odds ratios close to zero or infinity, or probabilities close to 0 or 1
- The main idea is that even in the absence of any data, we can usually judge some parameter values to be more realistic than others

Connection to Bayesian paradigm

- This is, of course, the central concept in Bayesian reasoning as well
- From a Bayesian perspective, $p(\boldsymbol{\theta})$ is simply the log-density of the prior distribution, and the objective function $q(\boldsymbol{\theta}|X)$ is the posterior log-density (up to a constant, and multiplied by -1)
- From this perspective, penalized likelihood simply means the study of the asymptotic/frequentist properties of the posterior mode, or MAP (*maximum a posteriori*) estimator

Theoretical considerations

- The theory of penalized likelihood is therefore governed, broadly speaking, by the Bernstein-von Mises theorem
- Recall that this theorem states that since $p(\boldsymbol{\theta})$ remains fixed as we collect an increasing amount of data, its contribution is negligible asymptotically and we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}^*))$$

  just as we would for the ordinary MLE
- If we trust this approximation, we may simply carry out inference using $\boldsymbol{\mathcal{I}}^{-1}$ or $\boldsymbol{\mathcal{J}}^{-1}$, but at the penalized MLE $\hat{\boldsymbol{\theta}}$

Variance estimation

- Alternatively, we can estimate standard errors by taking the second derivative of the objective function q:

$$\nabla^2 q = \nabla^2 \ell - \nabla^2 p$$

- If the likelihood is only lightly penalized, there is little difference between these two approaches
- If there is a modest amount of penalization, then typically this second approach (adjusting for the curvature of the likelihood by including $\nabla^2 p$) is more accurate

"Heavy" penalization

- If there is a heavy amount of penalization, however, then the asymptotic approximation is likely questionable in the first place
- Fundamentally, the likelihood will no longer be centered on the true parameter — even on average
  - Nonparametric smoothing, lasso, ridge, are all good examples of this and require a different inferential framework
  - It's not entirely clear from a philosophical standpoint what properties inference should even have in these scenarios (this is an ongoing area of research)
- Or you could carry out a fully Bayesian analysis

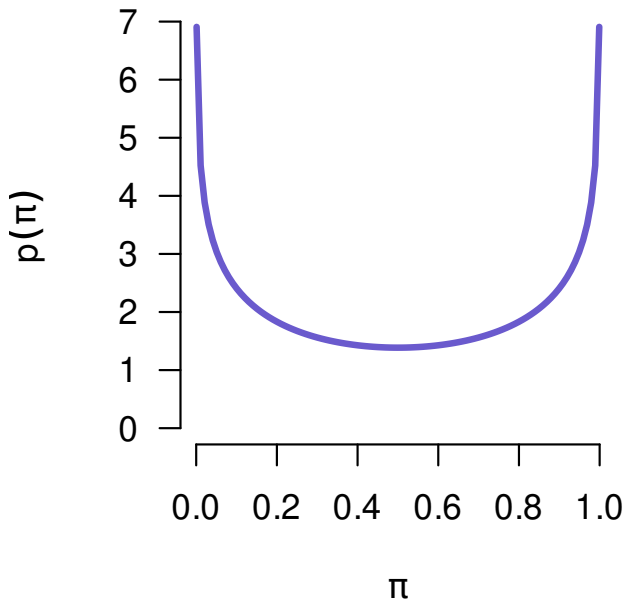
Binomial proportions

- Let's now see how this idea can be applied to a variety of problems, beginning with the simple problem of constructing confidence intervals for a binomial proportion
- As you are probably already aware, the Wald confidence interval is truly terrible in this situation
- The underlying problem is that the quadratic approximation breaks down near $\pi = 0$ and $\pi = 1$, where $\mathscr{I}(\pi) \to \infty$ and $\mathscr{I}^{-1}(\pi) \to 0$

Binomial proportions: Penalty

- Thus, it may be beneficial to penalize values of π near 0 and 1 (this is also just intuitively reasonable in most situations)
- This can be accomplished with the penalty function

$$-p(\pi) = \lambda \log \pi + \lambda \log(1 - \pi)$$



Resulting penalized likelihood

- From a Bayesian perspective, this penalty corresponds to the use of a Beta$(\lambda + 1, \lambda + 1)$ prior
- This results in the penalized likelihood

$$q(\pi) = (x + \lambda) \log \pi + (n - x + \lambda) \log(1 - \pi)$$

- Note that this is simply the likelihood of the original binomial distribution, but with λ successes and λ failures added to the observed data

Penalized Wald interval

- Using the penalized score and information, we have the penalized Wald interval $\hat{\pi} \pm z_{1-\alpha/2}\text{SE}$, where

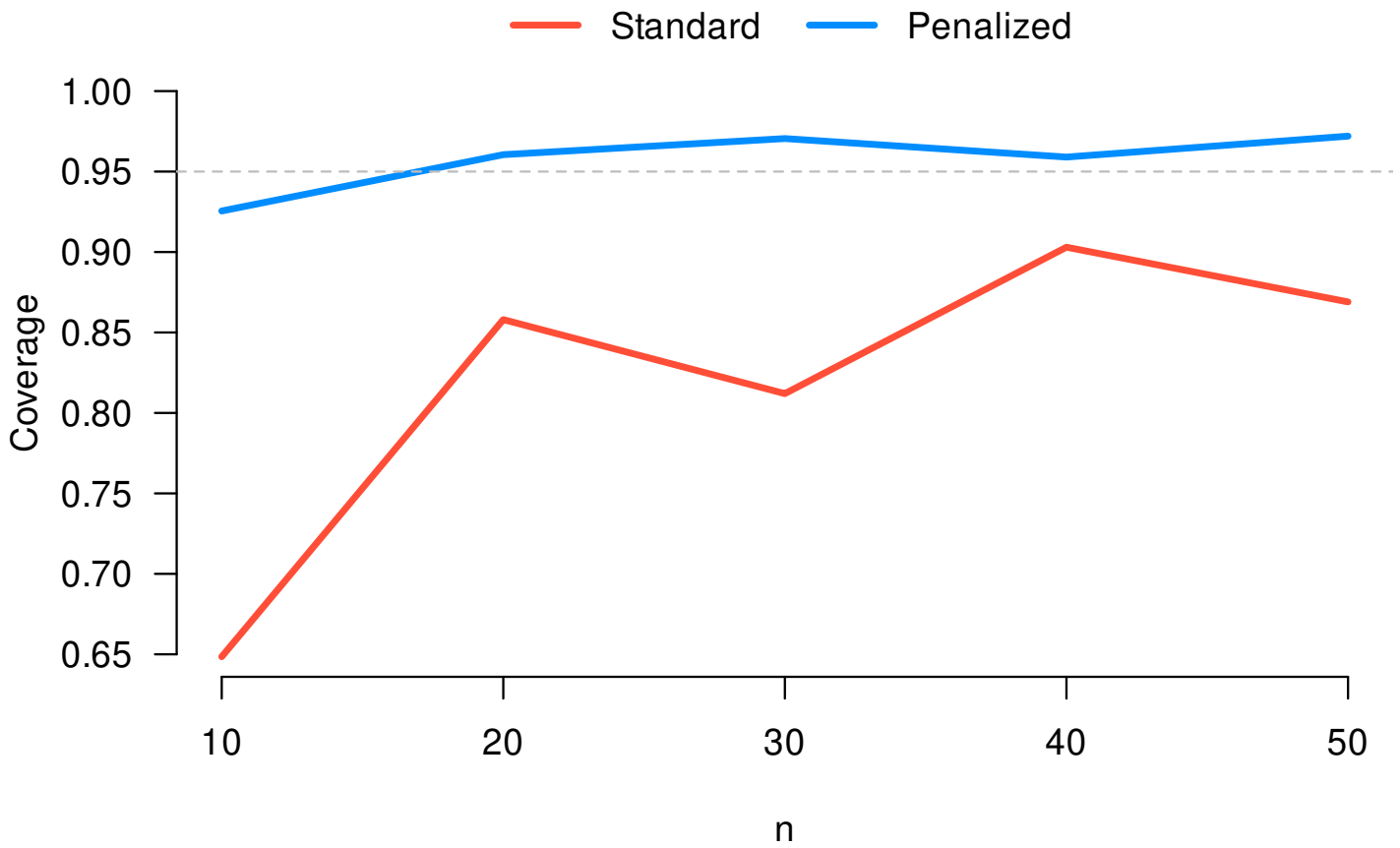
$$\hat{\pi} = \frac{x + \lambda}{n + 2\lambda}$$

$$\text{SE} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n + 2\lambda}}$$

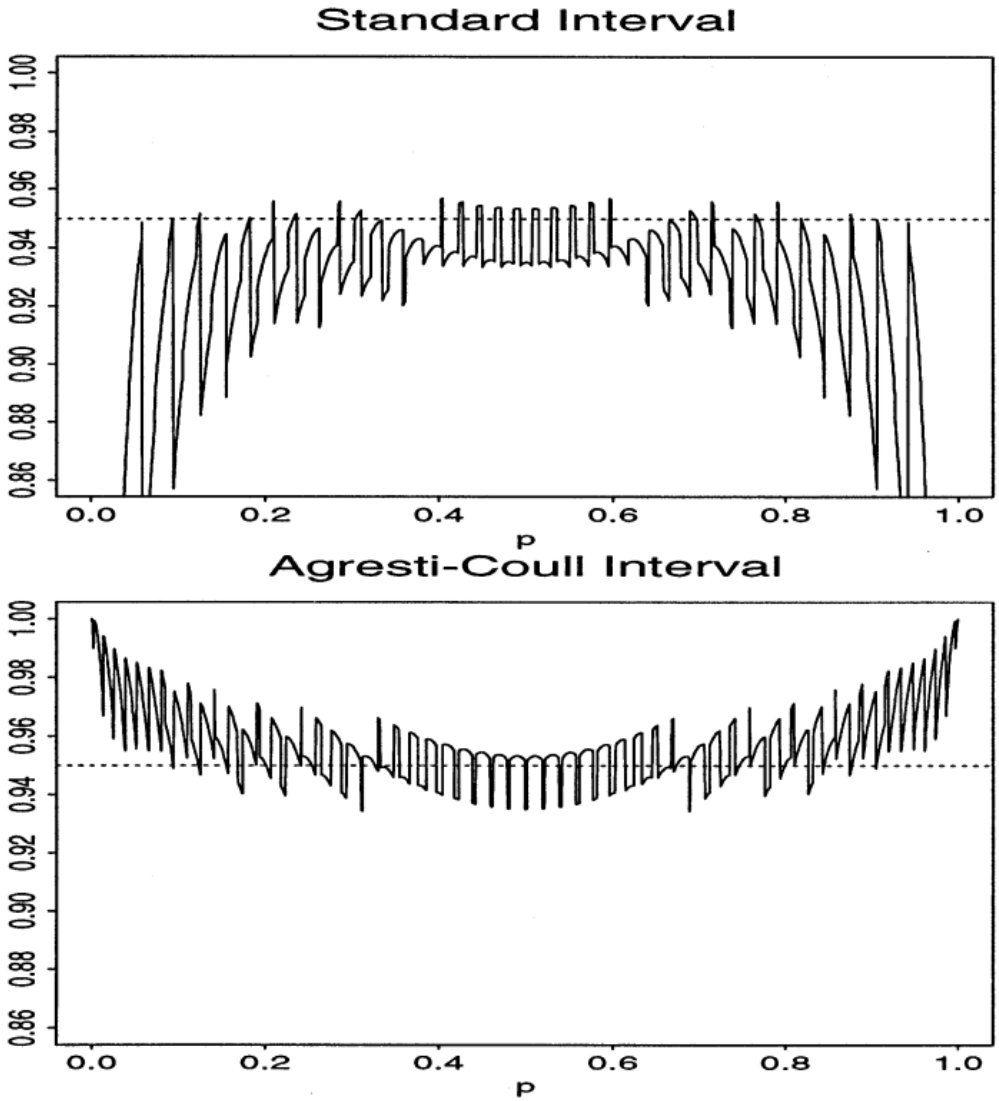
- The specific choice $\lambda = 2$ has been studied in the literature and is known as the Agresti-Coull interval

Simulation study

Empirical coverage of the two Wald intervals over 2,000 independent replications: $\pi^* = 0.1$, nominal coverage 95%



From Brown et al. (2001), $n = 50$



Complete separation

- Similar issues can arise in the regression setting
- In particular, when predictors are binary, it may happen that all observations where $X_{ij} = 1$ are cases; this is more likely to happen when n is small, when the number of predictors is large, when the event is rare, or when the predictor is rare
- When it does happen, the likelihood is no longer unimodal but instead increases without bound as $\beta_j \to \infty$
- This problem is known as *complete separation*

Complete separation: Example

- For example, suppose that we have 200 cases, 200 controls, that $X_1 \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, but that X_2 is binary and $x_{i2} = 1$ occurs only once
- In this situation, we will obtain results something like this (actual results are a bit arbitrary depending on the algorithm, since it cannot converge):

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	0.00	0.10	0.999
x1	-0.06	0.11	0.591
x2	22.50	48196.14	1.000

Firth penalized regression

- One approach to addressing this problem was proposed by David Firth in 1993
- His idea was to apply a penalty in the form of a Jeffreys prior to the logistic regression likelihood; this results in the penalized log-likelihood

$$\ell(\boldsymbol{\beta}) + \tfrac{1}{2}\log|\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})|,$$

  where $|\mathbf{A}|$ denotes the determinant of the matrix $\mathbf{A}$
- The resulting penalized score is

$$u_j^*(\boldsymbol{\beta}) = u_j(\boldsymbol{\beta}) + \tfrac{1}{2}\mathrm{tr}[\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\beta})\nabla_j\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})]$$

Firth penalized regression: Details

- Working through the linear algebra, we find that we can write the penalized score in terms of the "hat" matrix

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{1/2}$$

- Letting $\{h_i\}_{i=1}^n$ denote the diagonals of the hat matrix, we have

$$u_j^*(\boldsymbol{\beta}) = \mathbf{x}_j^\top(\mathbf{y} - \boldsymbol{\pi} + \mathbf{a}),$$

where $a_i = \frac{h_i}{2}(y_i - \pi_i + 1 - y_i - \pi_i)$

- Thus, we see a similar phenomenon as we saw earlier, where Firth regression is essentially adding a success and failure to the likelihood, here each with weight $h_i/2$

Returning to earlier example

- Applying this method to our earlier example with complete separation, we now have reasonable estimates and inference:

	Estimate	SE	p
(Intercept)	0.00	0.10	0.999
x1	-0.06	0.10	0.593
x2	1.03	1.64	0.531

- The standard error here is based on $\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, where $\mathbf{W}$ is evaluated at the penalized MLE
- This is typical in Firth regression, as the penalization is very slight

Remarks

- This penalization typically results in more accurate estimation, although its primary appeal is the elimination of complete separation; similar issues arise in Cox proportional hazards regression, to which Firth's idea is also often applied
- One can deal with complete separation in other ways, such as changing the model or dropping certain terms from the model if they introduce problems, but this is a bit unsatisfying and ad hoc, particularly if one is interested in rare risk factors
- Firth penalized regression is not as well known as it should be, but is widely used for example in genetic association studies, where logistic regression models are fit across thousands of genetic variants, some of which will be rare, and complete separation is virtually guaranteed to come up

Ridge regression

- The general idea of penalizing regression coefficients away from infinity can also be applied in linear regression
- One common way of doing so is to penalize the sum of the squares of the regression coefficients
- This idea, known as *ridge regression*, minimizes the objective function

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\beta_j^2$$

Ridge regression: Solution and information

- This yields the penalized MLE

$$\widehat{\boldsymbol{\beta}} = \sigma^{-2}(\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Furthermore, the penalized information matrix is

$$\boldsymbol{\mathcal{I}} = \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$$

- Note that compared with the standard information matrix, the penalized version has a "ridge" down the diagonal

Ridge regression: Unique solutions

- This ridge grants an important stability to ridge regression that ordinary linear regression lacks
- Namely, although the standard information matrix is not necessarily positive definite, the penalized information is (unless $\lambda = 0$)
- This confers many important benefits:
  - Solutions are always unique
  - Standard errors are always finite
  - Estimation much more accurate in the presence of correlated predictors

Ridge regression: Simple example

- To see how this is useful, let's just look a simple example with two highly correlated predictors:

```
# Ridge: Motivation
x1 <- rnorm(20)
x2 <- rnorm(20, mean = x1, sd = .01)
y <- rnorm(20, mean = 3 + x1 + x2)
coef(lm(y ~ x1 + x2))
# (Intercept)          x1          x2
#    3.021159   21.121729  -19.089170
```

- In this simulated example, we can directly observe that these estimates are wildly inaccurate
- However, even if we didn't know β^*, it is rather unlikely that X_1 has a large positive effect, but it just happens to be canceled out every time by a large effect from X_2

Ridge regression: Simple example (cont'd)

Let's see what ridge regression makes of this situation:

```
coef(lm.ridge(y ~ x1 + x2, lambda = 1))
#                   x1          x2
# 3.0489734 0.9874831 0.9585603
```

Much better than ordinary least squares

Additional uses

- Before ending this lecture, I'd like to briefly look at two other applications of penalized likelihood, as they pertain to smoothing and sparsity
- We don't have time to go into the details of how these methods work, but I think seeing the applications will be helpful to broaden your horizons in terms of thinking about situations in which penalization might be useful

Variable selection and sparsity

- We introduced ridge regression earlier and saw that it was very helpful for dealing with collinearity
- However, what if we wanted to impose a penalty that not only discouraged extremely large coefficient values, but also discouraged variables from even entering the model in the first place?
- We do this informally quite often in statistics when building models, iteratively adding and removing predictors from a model; could we do this automatically through the use of penalization?

Lasso

- Indeed we can, and with a surprisingly small change to ridge regression
- Suppose we penalize not the squared values of the coefficients, but their absolute values:

$$p(\boldsymbol{\beta}) = \sum_{j=1}^{d} |\beta_j|$$

- It turns out that doing so produces "sparse" penalized MLEs, in the sense that we have $\widehat{\beta}_j = 0$ exactly for some coefficients
- This method is called the lasso, for least absolute shrinkage and selection operator

Lasso: Toy example

- To see the general idea of how this works, let's apply it to a toy example:

```
# lasso
X <- matrix(rnorm(100 * 10), 100, 10)
y <- rnorm(100, 2 * X[, 1] + 3 * X[, 7])
fit <- glmnet(X, y)
coef(fit, s = 0.2)
#       s=0.2
# V1 1.633213
# V7 2.860543
```

  with all other coefficients zero
- If you are curious to learn more about this idea (and other penalized regression models), check out BIOS 7240: High dimensional data analysis

Nonparametric regression

- Consider the nonparametric regression problem in which $\mathbb{E}Y_i = f(x_i)$
- Suppose that we're willing to assume Y is normally distributed, but not willing to assume that its relationship with X is linear
- We would be interested in minimizing the residual sum of squares:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i))^2,$$

but this is clearly problematic as I could draw a wildly varying function that hits every y_i value

Illustration

Suppose, then, we introduce a penalty that discourages excessive "wiggliness":

$$p(f) = \lambda \int [f''(u)]^2 du$$