Ridge regression and mixed models

19 Oct 2023

Ridge in a nutshell

- **penalized** models: instead of minimizing $\text{SSQ} = \sum((\mathbf{y} - \mathbf{X}\beta)_i)^2$, minimize $\text{SSQ} + \lambda||\beta||_2$ (ridge)
- or + $||\beta||_1$ (lasso)
- optimize *bias-variance tradeoff*
- equivalent to imposing iid Gaussian priors on each element of β
- lasso (and elastic net, which is a convex combination of L2 and L1 penalties) are popular because they **induce sparsity**
  - *likelihood surfaces* are non-convex with cusps at zero
  - optimization with non-convex surfaces is a nuisance because it makes the basic optimization problem nonlinear; we need to use a different algorithm (coordinate descent/soft thresholding); can't use *only* linear algebra
- can generalize from penalized LM to penalized GLM

Andrew Gelman on variable selection

> Variable selection (that is, setting some coefficients to be exactly zero) can be useful for various reasons, including: * It's a simple form of regularization. * It can reduce costs in future data collection. Variable selection can be fine as a means to an end. Problems can arise if it's taken too seriously, for example as an attempt to discover a purported parsimonious true model.

Choosing penalty strength

- typically by *cross-validation*
- leave-one-out (LOOCV) vs k-fold

Practical points

- Predictors **must** be standardized
- Intercept should usually be unpenalized
- Avoid **data leakage**
  - don't include variables that are 'future' indicators of the outcome (e.g. see here)
  - full pipeline must be cross-validated (i.e. don't do data-dependent variable selection *before* cross-validating, or use the full data set to select a pipeline)
  - cross-validation must account for structure in the data
  - **either** ensure that residuals are *conditionally* independent
  - **or** take account of grouping structures in the data (block bootstrap, spatial stratification, etc. Wenger and Olden (2012)

Ridge vs lasso

- In practice people just try both (or elastic net)
- Conjecture: whether ridge or lasso is a better *predictive* model in a particular case depends on the *effect size spectrum*

Ridge by data augmentation

- set

$$\mathbf{B} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}$$

- and $\mathbf{y}^* = (\mathbf{y}\ \mathbf{0})$
- so that $\mathbf{B}^\top \mathbf{B} = \mathbf{X}^\top \mathbf{X} + \lambda I$ and the residual sum of squares is unchanged

Inference

- inference from penalized models is really hard
- classical CIs for ridge are **identical** to OLS (Obenchain 1977) > ridge techniques do not generally yield `new'' normal theory statistical inferences: in particular, ridging does not necessarily produceshifted'' confidence regions.
- **no free lunch** (i.e., no true narrowing of CIs/decreased uncertainty without additional assumptions)
- post-selection inference is a big deal but requires very strong assumptions (asymptotic, 'gap')
- prediction intervals are often neglected (conformal prediction, jackknife+ (Barber et al. 2021)): MAPIE

Practical

- `glmnet` is very good
- `ridge`, `lmridge`, … ('library(sos); findFn("{ridge regression}")1)
- need to give `y` and `X` directly (although see glmnetUtils package)

Tangent: how do I know if an R package is any good?

- how old is it/how many releases has it had?
- is it actively developed?
- does the documentation give literature citations?
- does it have reverse dependencies?
- what is its ranking on CRAN? `packageRank::packageRank("lmridge")` (80th percentile)

James-Stein estimators

- more formally, why is ridge better?
- based on a single observation, $\mathbf{y}$, of a *multivariate* response with dimension $m \geq 3$, shrinking the value (usually toward zero) is a better estimate of the mean than the value itself

From ridge to mixed models

- what if we say

$$
\begin{aligned}
\mathbf{y} &\sim \text{Normal}(\mathbf{X}\beta, \sigma^2) \\
\beta &\sim \text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{I})
\end{aligned}
$$

?

i.e. treat this as an *empirical Bayesian* problem (we estimate the β values, but do not put a prior on σ^2 or a hyperprior on σ_g^2 $(= 1/\lambda)$

References

Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. "Predictive Inference with the Jackknife+." *The Annals of Statistics* 49 (1): 486–507. https://doi.org/10.1214/20-AOS1965.

Obenchain, R. L. 1977. "Classical F-Tests and Confidence Regions for Ridge Regression." *Technometrics* 19 (4): 429–39. https://doi.org/10.1080/00401706.1977.10489582.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8): 913–29. https://doi.org/10.1111/ecog.02881.

Wenger, Seth J., and Julian D. Olden. 2012. "Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation." *Methods in Ecology and Evolution* 3 (2): 260–67. https://doi.org/10.1111/j.2041-210X.2011.00170.x.