Ridge regression and mixed models

3 Sep 2024

Ridge in a nutshell

- **penalized** models: instead of minimizing SSQ $= \sum((\mathbf{y}-\mathbf{X}\beta)_i)^2$, minimize 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)

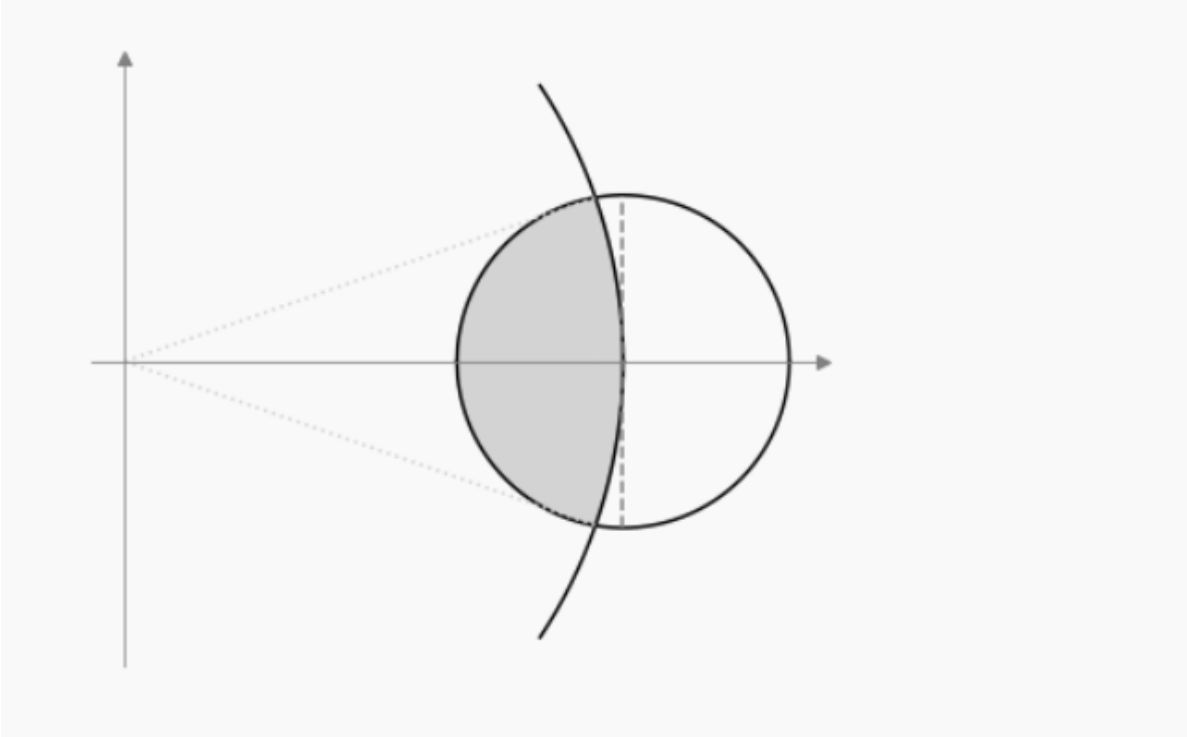
James-Stein estimators

- 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 (!)

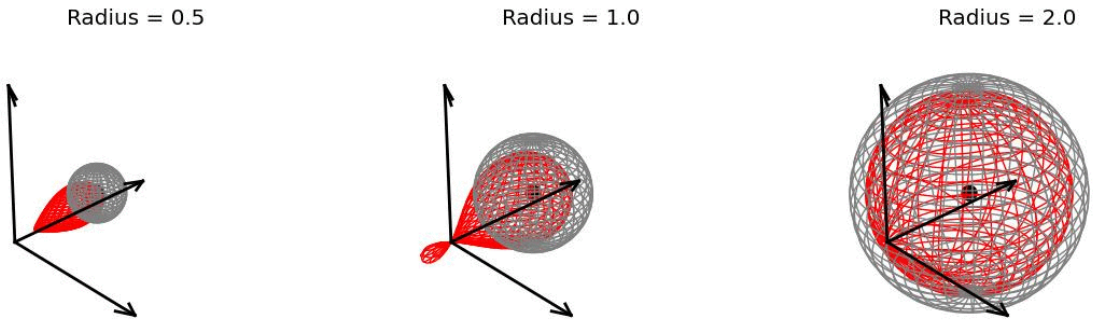
$$\hat{\mu}(X_1, \ldots, X_n) = \left(1 - \frac{(p-2)\sigma^2/n}{||\bar{X}_n||^2}\right) \bar{X}_n$$

- (connected to recurrence of random walks in $d \leq 2$, non-recurrence in $d \geq 3$ ...)
- "paradox": the quantities in the vector don't have to have anything to do with each other (and, we can shrink to any point, not necessarily zero ...)

From Antognini (2021):



From Harris (2013):



- van Houwelingen (2001) gives a very nice explanation/transition from James-Stein to penalized regression etc.

From ridge to mixed models

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 $(= 1/\lambda)$

From van Houwelingen (2001) (ultimately from Efron and Morris 1972):

If we use a prior with $\mu_i \sim N(\mu, \tau^2)$ (assuming residual variance is 1 wlog), then

$$E(\mu_i|X_i) = \mu + \frac{\tau^2}{\tau^2+1}(X_i - \mu)$$
$$var(\mu_i|X_i) = \frac{\tau^2}{\tau^2+1}$$

But we still have to estimate τ (or $\tau^2/(\tau^2+1)$ from the data).

MVN version

We can be much more general:

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

Back to 1D

The simplest case (described in an R formula as `y ~ 1 + (1|g)`) is a model with a population-level intercept β_0 and group-level deviations from the population mean b_i.

This case, and more complex cases, can be written as

$$y_i \sim \text{Normal}((\mathbf{X}\beta + \mathbf{Zb})_i, \sigma_r^2)$$
$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \Sigma(\theta))$$

where θ is a vector of parameters that defines the covariance matrix Σ.

How do we estimate this?

- can use EM algorithm (e.g. see here, or the lmm package)
- Or by linear algebra. For LMMs, we do a more complicated version of *data augmentation*.
- given a value for the random-effects variance, we can calculate the log-likelihood in one step (see
- large, sparse matrix computation
- has to be done *repeatedly*
- most efficient if we analyze the matrix and permute to optimize structure (Bates et al. 2015)
- then we need to do some kind of search over the space of variances
- derivatives are available in particular special cases

The general case

Given a model of the form

$$
\begin{aligned}
y_i &\sim \text{Normal}((\mathbf{X}\beta + \mathbf{Zb})_i, \sigma_r^2) \\
\mathbf{b} &\sim \text{MVN}(\mathbf{0}, \Sigma(\theta))
\end{aligned}
$$

- How do we specify and set up Z?
- How do we specify and set up Σ?

constructing the random-effects model matrix

- specify as `(t|g)`; `t` is the *term* and `g` is the *grouping factor*
- for intercepts, just the indicator matrix
- for more complex models (random slopes), take the *Khatri-Rao* product of the model matrix of the term with the indicator matrix of `g`
- concatenate multiple random effects terms into a single `Z` matrix

constructing the covariance matrix

- blockwise
- what's the best way to parameterize a positive-(semi)definite matrix? (Pinheiro and Bates 1996)
- Cholesky decomposition with
- scaled or unscaled?
- Cholesky or log-Cholesky scale?

- separating correlation and SD vectors: glmmTMB:

$$\Sigma = D^{-1/2} L L^\top D^{-1/2}, \quad D = \text{diag}(L L^\top)$$

References

Antognini, Joe. 2021. "Understanding Stein's Paradox." https://joe-antognini.github.io/machine-learning/steins-paradox.

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.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67 (October): 1–48. https://doi.org/10.18637/jss.v067.i01.

Harris, Naftali. 2013. "Visualizing the James-Stein Estimator." https://www.naftaliharris.com/blog/steinviz/.

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.

Pinheiro, José C., and Douglas M. Bates. 1996. "Unconstrained Parametrizations for Variance-Covariance Matrices." *Statistics and Computing* 6 (3): 289–96. https://doi.org/10.1007/BF00140873.

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.

van Houwelingen, J. C. 2001. "Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy." *Statistica Neerlandica* 55 (1): 17–34. https://doi.org/10.1111/1467-9574.00154.

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.