

Ridge regression/Tikhonov regularization

Object function and estimates. In ridge regression (Tikhonov regularization), one estimates the parameter vector β of the linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad (1)$$

by minimization of the object function

$$\chi^2 = \|\mathbf{X}\beta - \mathbf{y}\|^2 + h^2\|\beta\|^2. \quad (2)$$

The first term of the object function is the residual sum of squares; the second term penalizes a large norm (“roughness”) of the parameter vector β . The regularization parameter (ridge parameter) h determines the trade-off between minimizing the residual sum of squares and minimizing the norm of the estimate. For a given regularization parameter h , the regularized estimate is

$$\hat{\beta}_h = (\mathbf{X}^T\mathbf{X} + h^2\mathbf{1})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3)$$

For regularization parameter $h = 0$, the estimate $\hat{\beta}_h$ reduces to the ordinary least squares estimate. For regularization parameter $h \rightarrow \infty$, the estimate $\hat{\beta}_h$ approaches zero. For intermediate values of the regularization parameter, the estimate $\hat{\beta}_h$ is “shrunk” toward zero compared with the ordinary least squares estimate; hence, it is a biased estimate. Even if the design matrix \mathbf{X} is rank-deficient, so that $\mathbf{X}^T\mathbf{X}$ is singular, the regularized matrix $(\mathbf{X}^T\mathbf{X} + h^2\mathbf{1})$ is nonsingular for any nonzero value of h .

Estimates in terms of SVD. Substituting the singular value decomposition of the $n \times p$ design matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ into the ridge regression estimate (3) yields

$$\hat{\beta}_h = \sum_{i=1}^p f_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i \quad (4)$$

with *filter factors*

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + h^2}. \quad (5)$$

The estimate $\hat{\beta}_h$ is a linear combination of right singular vectors \mathbf{v}_i of the design matrix \mathbf{X} . For singular values $\sigma_i \ll h$, the filter factors f_i are approximately

zero and the associated right singular vectors \mathbf{v}_i are filtered out. For singular values $\sigma_i \gg h$, the filter factors f_i are approximately one and the associated right singular vectors \mathbf{v}_i are retained in the estimate $\hat{\boldsymbol{\beta}}_h$.

In signal processing, the filter function (5) is known as a *Wiener filter*. The Wiener filter, acting on Fourier coefficients, is an optimal filter to separate a signal from noise. The squared singular values σ_i^2 are the correlate of the spectral density of the signal in Wiener filtering. The squared regularization parameter h^2 is the correlate of the spectral density of the noise in Wiener filtering (cf. Papoulis 1991, chapter 14.1). In ridge regression, right singular vectors \mathbf{v}_i for which the ratio of “signal” σ_i^2 to “noise” h^2 is much smaller than one are filtered out.

Filtering and comparison with TSVD. The estimate of the parameter vector $\boldsymbol{\beta}$ regularized by truncated SVD can be written in the same form as the ridge regression estimate (3),

$$\hat{\boldsymbol{\beta}}_{\hat{r}} = \sum_{i=1}^p f_i \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i, \quad (6)$$

but with filter factors

$$f_i = \begin{cases} 1 & \text{for } i \leq \hat{r} \\ 0 & \text{for } i > \hat{r} \end{cases} \quad (7)$$

(for truncation at effective rank \hat{r}). That is, regularization by TSVD corresponds to filtering with a step function filter.

The ordinary least squares estimate can likewise be written in the form (6), with filter factors $f_i = 1$ for all i (that is, no filtering).

In TSVD regularization, the regularization parameter \hat{r} is discrete; in ridge regression, the regularization parameter h is continuous. The filter factors f_i in ridge regression decay more slowly with decreasing singular values σ_i than the filter factors in TSVD regularization. Ridge regression affords a smoother filtering than TSVD.

The structural parallels between the ridge regression filter (5) and the optimal Wiener filter suggest that ridge regression might suppress noise in data in a more robust way and with less loss of relevant information than TSVD. Indeed, ridge regression also arises as a regularization method when errors in the design matrix \mathbf{X} , which are ignored in the regression model (1), are taken into account (Golub et al. 2000). However, since the adequacy of regularization methods is context-dependent (e.g., in image processing, smoothness of the estimates is not always desirable), it is difficult to make general statements about the superiority of one regularization method over another.

Transformation to standard form. The object function (2) is said to be in standard form. In place of the standard-form object function (2), one often wants to minimize an object function of the form

$$\chi^2 = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + h^2\|\mathbf{L}\boldsymbol{\beta}\|^2. \quad (8)$$

If smoothness of the estimate is an objective, the weight matrix \mathbf{L} might represent a discrete second-derivative operator. [Discretizing the second derivative of a function f as $f''_i \approx \Delta^{-2}(f_{i-1} - 2f_i + f_{i+1})$, with step size Δ , the matrix \mathbf{L} would be a tridiagonal matrix with $-2\Delta^{-2}$ on the main diagonal and Δ^{-2} on the adjacent diagonals (except for boundary points).] Another common choice of weight matrix \mathbf{L} is a diagonal matrix with diagonal elements $L_{kk} = \sqrt{(\mathbf{X}^T\mathbf{X})_{kk}}$. With this choice of weight matrix, the estimated parameter vector $\hat{\boldsymbol{\beta}}_h$ is invariant under rescaling of the response vector \mathbf{y} .

If the weight matrix \mathbf{L} is invertible, the object function (8) can be brought into standard form by the transformation $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{L}^{-1}$ and $\tilde{\boldsymbol{\beta}} = \mathbf{L}\boldsymbol{\beta}$. The standard-form ridge regression can then be computed from the rescaled design matrix $\tilde{\mathbf{X}}$. If the weight matrix \mathbf{L} is not invertible, other methods must be used (see Hansen 1998, chapter 2.3).

REFERENCES

- Golub, G. H., P. C. Hansen, and D. P. O’Leary, 2000: Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, **21**, 185–194.
- Hansen, P. C., 1998: *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM Monogr. on Mathematical Modeling and Computation, Society for Industrial and Applied Mathematics. [An excellent research monograph on regularization methods.]
- Papoulis, A., 1991: *Probability, Random Variables, and Stochastic Processes*. 3rd ed., McGraw Hill.