The Generalized Ridge Trace Plot: Visualizing Bias and Precision

Michael Friendly
Psychology Department and Statistical Consulting Service
York University
4700 Keele Street, Toronto, ON, Canada M3J 1P3

In press: *Journal of Computational and Graphical Statistics*

BibTeX:

```
@Article {Friendly:genridge:2012,
   author = {Friendly, Michael},
   title = {The Generalized Ridge Trace Plot: Visualizing Bias and Precision},
   journal = {Journal of Computational and Graphical Statistics},
   year = {2012},
   volume = {},
   number = {},
   note = {In press; Accepted, 2/1/2012},
   url = {http://datavis.ca/papers/genridge.pdf}
}
```

The Generalized Ridge Trace Plot: Visualizing Bias *and* Precision

Michael Friendly
York University, Toronto

Rev. 3, February 2, 2012

Abstract

In ridge regression and related shrinkage methods, the ridge trace plot, a plot of estimated coefficients against a shrinkage parameter, is a common graphical adjunct to help determine a favorable tradeoff of bias against precision (inverse variance) of the estimates. However, standard unidimensional versions of this plot are ill-suited for this purpose because they show only bias directly and ignore the multi-dimensional nature of the problem.

We introduce a generalized version of the ridge trace plot, showing covariance ellipsoids in parameter space, whose centers show bias and whose size and shape show variance and covariance in relation to the criteria for which these methods were developed. These provide a direct visualization of *both* bias and precision. Even 2D bivariate versions of this plot show interesting features not revealed in the standard univariate version. Low-rank versions of this plot, based on an orthogonal transformation of predictor space extend these ideas to larger numbers of predictor variables, by focusing on the dimensions in the space of predictors that are likely to be most informative about the nature of bias and precision. Two well-known data sets are used to illustrate these graphical methods. The `genridge` package for R implements computation and display.

Key words: biplot; model selection; multivariate bootstrap; regression shrinkage; ridge regression; ridge trace plot; singular value decomposition; variance-shrinkage tradeoff

1 Introduction

We consider the classical linear model for a univariate response, $\boldsymbol{y} = \beta_0 \boldsymbol{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, $\mathsf{Var}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}) = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{X}$ is $(n \times p)$ and of full rank. In this context, high multiple correlations among the predictors lead to well-known problems of collinearity under ordinary least squares (OLS) estimation, which result in unstable estimates of the parameters in $\boldsymbol{\beta}$: standard errors are inflated and estimated coefficients tend to be too large in absolute value on average.

Ridge regression and related shrinkage methods have a long history, initially stemming from problems associated with OLS regression with correlated predictors (Hoerl and Kennard, 1970a) and more recently encompassing a wide class of model selection methods, of which the LASSO method of Tibshirani (1996) and LAR method of Efron *et al.* (2004) are well-known instances. See, for example, the reviews in Vinod (1978) and McDonald (2009) for details and context omitted here.

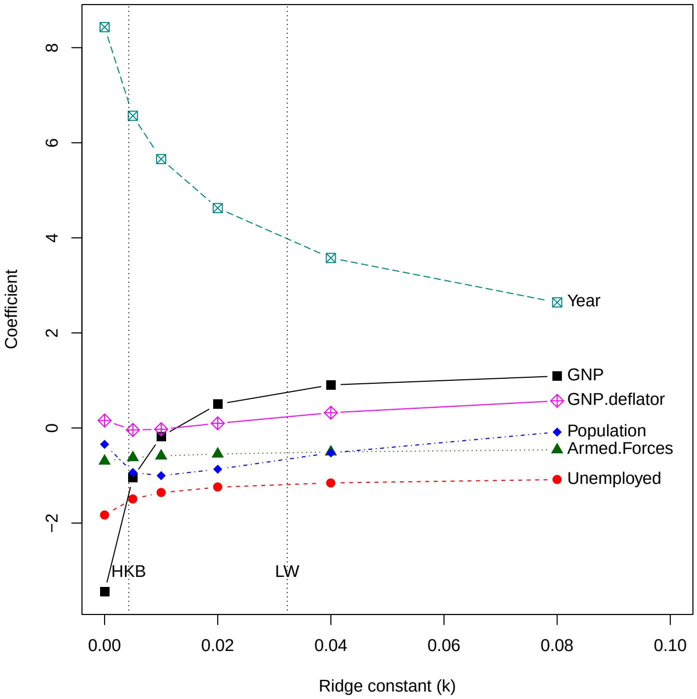

Figure 1: Univariate ridge trace plots for the coefficients of predictors of Employment in Longley's data via ridge regression, with ridge constants $k = 0, 0.005, 0.01, 0.02, 0.04, 0.08$. The dotted lines show choices for the ridge constant by two commonly-used criteria due to HKB: Hoerl *et al.* (1975) and LW: Lawless and Wang (1976). What can you see here to decide about the tradeoff of bias against precision?

An essential idea behind these methods is that the OLS estimates are constrained in some way, shrinking them, on average, toward zero, in order to satisfy other considerations. In the modern literature on model selection, interest is often focused on predictive accuracy, since the OLS estimates will typically have low bias but large prediction variance. The general goal of these methods is to achieve a more favorable trade-off between bias and variance to improve overall predictive accuracy.

Another common characteristic of these methods is that they involve some tuning parameter (k) or criterion to quantify the tradeoff between bias and variance. In many cases, analytical or computationally intensive methods have been developed to choose an optimal value of the tuning parameter, for example using generalized cross validation, bootstrap methods, or any of a number of criteria developed through extensive simulation studies (Gibbons, 1981).

Our interest here concerns the graphical methods that are commonly used as adjuncts to these methods, to display how the estimated coefficients are affected by the model shrinkage or selection method, and perhaps to allow the analyst to adjust the tuning parameter or criterion to take account of substantive or other non-statistical considerations. The prototype of such graphical displays is the (univariate) ridge-trace plot, introduced by Hoerl and Kennard (1970b). An illustration of this graphical form is shown in Figure 1, using an example of ridge regression described in Section 3.

It is commonly believed (McDonald, 2009) that such plots provide a visual assessment of the effect of the choice of k that supplements the multitude of (often diverging) numerical criteria, thus allowing the analyst to make more informed decisions.

For interpretative purposes, such plots are sometimes annotated with vertical lines showing the tuning constant selected by one or more methods (as in Figure 1), or worse, are superposed with separate graphs showing some measure of predictive variance, of necessity on a separately scaled vertical axis, thereby committing a venal, if not cardinal, graphical sin. These graphs fail their intended purpose—to display the tradeoff between bias and variance—because they use the wrong graphic form: an essentially univariate plot of trace lines for what is essentially a multivariate problem.

In this article, I describe and illustrate a multivariate generalization of the ridge-trace plot, based on consideration of the primary quantities to be estimated: a p-vector of estimated coefficients, $\boldsymbol{\beta}^{\star}(k)$ and its associated variance-covariance matrix, $\mathsf{Var}[\boldsymbol{\beta}^{\star}(k)]$, as a function of some tuning constant, k. I do not assert that the univariate ridge-trace plot (of $\boldsymbol{\beta}^{\star}(k)$ vs k) has no value, but rather she cannot escape flatland, where her higher-dimensional cousins have more to offer.

To give the flavor of this generalized ridge-trace plot, Figure 2 shows one version for the Longley data, discussed in more detail in Section 3.1. What you should see here is that, for pairs of coefficients, the estimated coefficients (centers of the ellipses) are driven in a systematic path as the ridge constant k varies. This gives additional information about how the change in one coefficient can affect other coefficients. Second, the relative size and orientation of the ellipses, corresponding to bivariate confidence regions, show directly the relative precision associated with each ridge estimate.

Because my focus is only on this graphical extension of ridge-trace plots, this introduction has been kept brief, mainly conceptual, and I have omitted all but a few key references. In particular, I sidestep critical discussion of deeper issues concerning the appropriate use and interpretation of these shrinkage methods and take liberties with the term "bias," which strictly speaking requires consideration of the true but unknown parameters. In what follows, I also restrict attention largely to the context of ridge regression. However, there is no loss of generality here, because the graphical ideas apply to *any* method that yields a set of estimates $\boldsymbol{\beta}^{\star}(k)$ and their covariance matrices, $\mathsf{Var}[\boldsymbol{\beta}^{\star}(k)]$, indexed by a set of tuning constants, $\boldsymbol{k}$.

1.1 Properties of Ridge Regression

To provide context, notation and some useful results, we provide a capsule summary of ridge regression here. So as to avoid unnecessary details related to the intercept, assume the predictors have been centered at their means and the unit vector is omitted from $\boldsymbol{X}$. Further, to avoid scaling issues, we rescale the columns of $\boldsymbol{X}$ to unit length, so that $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is a correlation matrix. Then, the OLS estimates are given by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \ , \tag{1}$$

with $\widehat{\mathsf{Var}}(\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}) = \widehat{\sigma}^2(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}$. Ridge regression replaces the standard residual sum of squares criterion with a penalized form,

$$\mathrm{RSS}(k) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} \qquad (k \geq 0) \ , \tag{2}$$

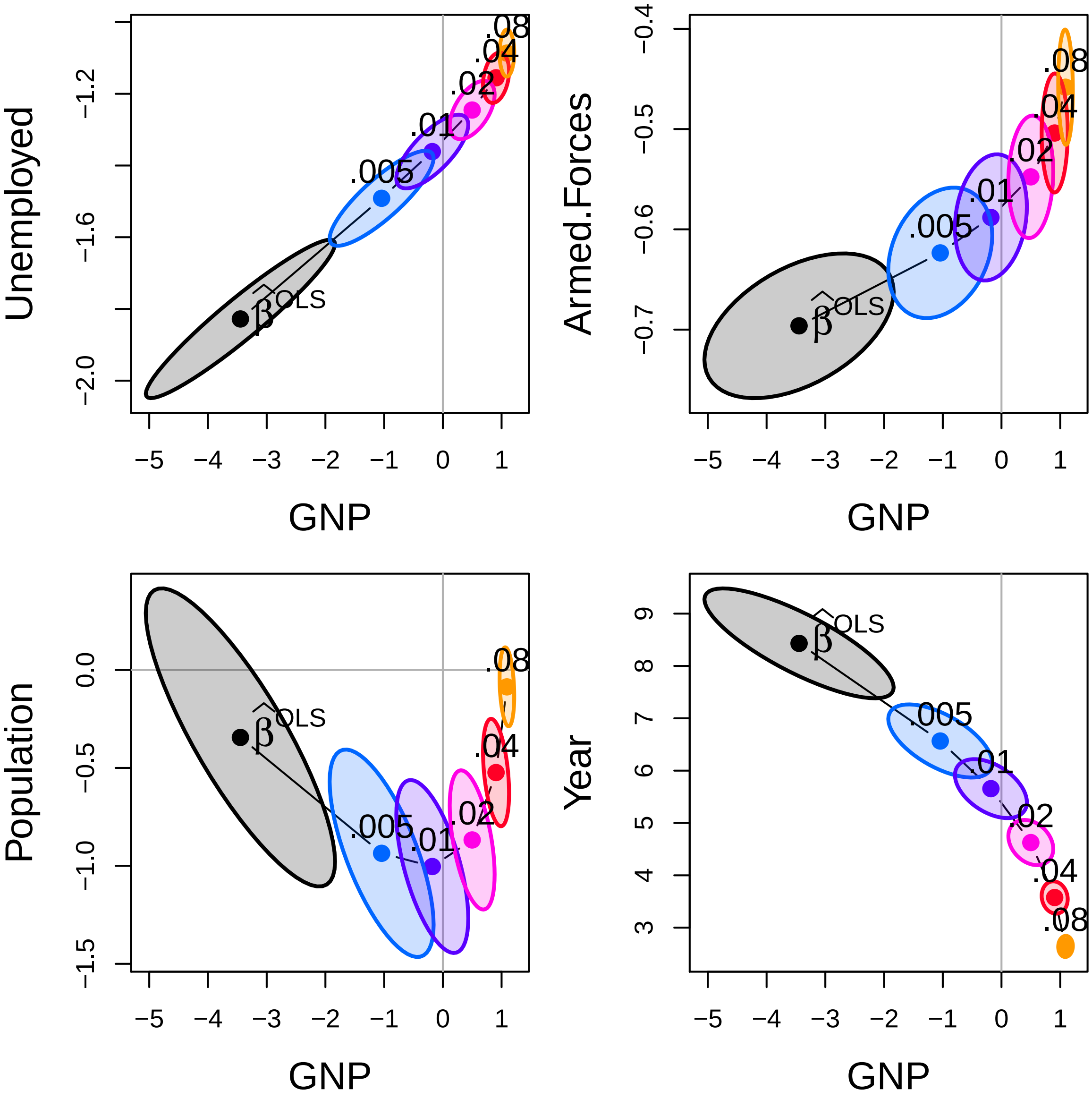

Figure 2: Bivariate ridge trace plots for the coefficients of four predictors against the coefficient for GNP in Longley's data, with $k = 0, 0.005, 0.01, 0.02, 0.04, 0.08$. (Corresponding values of df_k range from 6 to 4.09.) In most cases the coefficients are driven toward zero, but the bivariate plot also makes clear the reduction in variance, as well as the bivariate path of shrinkage. To reduce overlap, all covariance ellipses are shown with 1/2 the standard unit radius.

whose solution is easily seen to be

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{\text{RR}}_k &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ &= \boldsymbol{G}_k\,\widehat{\boldsymbol{\beta}}^{\text{OLS}}\ , \end{aligned} \tag{3}$$

where $\boldsymbol{G}_k = \left[\boldsymbol{I} + k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right]^{-1}$. Thus, as k increases, $\boldsymbol{G}_k$ decreases, driving $\widehat{\boldsymbol{\beta}}^{\text{RR}}_k$ toward $\boldsymbol{0}$ (Hoerl and Kennard, 1970a,b). The addition of a positive constant k to the diagonal of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ drives $|\boldsymbol{X}^\mathsf{T}\boldsymbol{X} +$

$k\boldsymbol{I}|$ away from zero even if $|\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}| \approx 0$. The estimated variance-covariance of $\widehat{\boldsymbol{\beta}}_k^{\mathrm{RR}}$ can then be expressed as

$$\widehat{\mathsf{Var}}(\widehat{\boldsymbol{\beta}}_k^{\mathrm{RR}}) = \widehat{\sigma}^2 \boldsymbol{G}_k (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1} \boldsymbol{G}_k^{\mathsf{T}} \ . \tag{4}$$

Eqn. (3) is computationally expensive, potentially numerically unstable for small k, and conceptually opaque, in that it sheds little light on the underlying geometry of the data in the column space of $\boldsymbol{X}$. An alternative formulation can be given in terms of the singular value decomposition (SVD) of $\boldsymbol{X}$,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}} \tag{5}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are respectively $n \times p$ and $p \times p$ orthonormal matrices, so that $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}$, and $\boldsymbol{D} = \mathrm{diag}\,(d_1, d_2, \dots d_p)$ is the diagonal matrix of ordered singular values, with entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Since $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathsf{T}}$, the eigenvalues of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ are given by $\boldsymbol{D}^2$ and therefore the eigenvalues of $\boldsymbol{G}_k$ can be shown (Hoerl and Kennard, 1970a) to be the diagonal elements of

$$\boldsymbol{D}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D} = \mathrm{diag}\left(\frac{d_i^2}{d_i^2 + k}\right) \ . \tag{6}$$

Noting that the eigenvectors, $\boldsymbol{V}$ are the principal component vectors, and that $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}$, the ridge estimates can be calculated more simply in terms of $\boldsymbol{U}$ and $\boldsymbol{D}$ as

$$\widehat{\boldsymbol{\beta}}_k^{\mathrm{RR}} = (\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} = \left(\frac{d_i}{d_i^2 + k}\right) \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y}, \quad i = 1, \dots p \tag{7}$$

and the fitted values can be expressed as

$$\begin{aligned} \widehat{\boldsymbol{y}}_k^{\mathrm{RR}} &= \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \\ &= \sum_i^p \boldsymbol{u}_i \left(\frac{d_i^2}{d_i^2 + k}\right) \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y} \end{aligned}$$

The terms $d_i^2/(d_i^2 + k) \leq 1$ are thus the factors by which the coordinates of $\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y}$ are shrunk with respect to the orthonormal basis for the column space of $\boldsymbol{X}$. The small singular values d_i correspond to the directions which ridge regression shrinks the most. These are the directions which contribute most to collinearity, for which other visualization methods have been proposed (Friendly and Kwan, 2009).

This analysis also provides an alternative and more intuitive characterization of the ridge tuning constant. By analogy with OLS, where the hat matrix, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}$ reflects degrees of freedom $\mathrm{df} = \mathrm{tr}(\boldsymbol{H}) = p$ corresponding to the p parameters, the effective degrees of freedom for ridge regression (Hastie *et al.*, 2001) is

$$\begin{aligned} \mathrm{df}_k &= \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathsf{T}}] \\ &= \sum_i^p \mathrm{df}_k(i) = \sum_i^p \left(\frac{d_i^2}{d_i^2 + k}\right) \ . \end{aligned} \tag{8}$$

Eqn. (8) is a monotone decreasing function of k, and hence any set of ridge constants can be specified in terms of equivalent df_k.

We note here the close connection with principal components regression. Ridge regression shrinks all dimensions in proportion to $\text{df}_k(i)$, so the low variance dimensions are shrunk more. Principal components regression discards the low variance dimensions and leaves the high variance dimensions unchanged.

2 Generalized Ridge Trace Plots: Theory and Methods

The essential idea is very simple: Rather than just plotting the univariate trajectories of the estimated coefficients vs. k, we plot the covariance ellipsoids of the estimated parameters over the same range of k.

The centers of these ellipsoids, for each predictor, provide the same information as in the standard univariate trajectories. In addition, in bivariate and multivariate views, they provide information about how the change in the ridge estimate for one parameter is related to simultaneous changes in the estimates for *other* parameters as a result of the shrinkage imposed by k (or df_k).

Moreover, the size and shape of the covariance ellipsoids show *directly* the effect on precision of the estimates as a function of k. For example, in bivariate views, we can see how the ridge constant affects the estimated standard errors for both variables, as well as the covariance of those parameters.

2.1 Details: "Confidence" ellipsoids

Specifically, assume that we have a series of estimates of the coefficients, $\widehat{\boldsymbol{\beta}}^{\star}_k$, each with an associated estimated covariance matrix, $\widehat{\boldsymbol{\Sigma}}^{\star}_k \equiv \widehat{\mathsf{Var}}(\widehat{\boldsymbol{\beta}}^{\star}_k)$, for example as in Eqn. (3) and Eqn. (4). Then the generalized ridge trace plot is a graphical representation of the set of the covariance ellipsoids $\mathcal{E}(\widehat{\boldsymbol{\beta}}^{\star}_k, \widehat{\boldsymbol{\Sigma}}^{\star}_k)$, where the envelope of an ellipsoid of radius c is defined by

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) = c^2\} \ . \tag{9}$$

In the context of hypothesis tests and confidence regions for parameters under classical, normal theory, the radius $c = \sqrt{dF_{1-\alpha}(d, \text{df}_e)}$ will give confidence ellipsoids of coverage $1 - \alpha$ for tests or regions of dimensionality d, with df_e degrees of freedom for $\boldsymbol{\Sigma}$. We don't use this here and will generally take c to be some convenient constant to give a reasonable separation among the ellipsoids for varying values of k. All that matters for the present purposes is that the same value of c is used for all ellipsoids in a given plot. Thus, the covariance ellipsoids we show here are meant to be interpreted here only in terms of their relative sizes and shapes, over a range of k, but not as strict confidence ellipsoids with any given coverage.

A computational definition of an ellipsoid that corresponds to the above in the case of positive-definite matrices $\boldsymbol{\Sigma}$ is

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu} \oplus \boldsymbol{A}\mathcal{S} \ , \tag{10}$$

where $\mathcal{S}$ is a unit sphere of conformable dimension and $\boldsymbol{\mu}$ is the centroid of the ellipsoid and where the $\oplus$ operator translates an ellipsoid to a given centroid. One convenient choice of $\boldsymbol{A}$ is the Choleski square root, $\boldsymbol{\Sigma}^{1/2}$. Eliding the unit sphere, we can express $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu} \oplus \boldsymbol{\Sigma}^{1/2}$. For details on the theory, computation and applications of ellipsoids in statistical practice, see Friendly *et al.* (2011).

2.2 Bootstrap methods

If normal theory on which these classical ellipsoids is deemed too restrictive, or if there is no closed-form expression for the variance-covariance matrix for some other shrinkage method, simple non-parametric versions can be calculated via bootstrap methods as follows: Generate B bootstrap estimates $\widetilde{\boldsymbol{\beta}}_k^b, b = 1, 2, \ldots, B$ by resampling from the rows of available data, $(\boldsymbol{y}, \boldsymbol{X})$. For given k, the bootstrap estimate $\widetilde{\boldsymbol{\beta}}_k$ is then the average over bootstrap samples and the bootstrap estimate $\widetilde{\boldsymbol{\Sigma}}_k$ of the covariance matrix of parameters can then be computed as the empirical covariance matrix of $\widetilde{\boldsymbol{\beta}}_k^b$ across the bootstrap samples,

$$\widetilde{\boldsymbol{\Sigma}}_k = B^{-1} \sum_{b=1}^{B} (\widetilde{\boldsymbol{\beta}}_k^b - \widetilde{\boldsymbol{\beta}}_k)(\widetilde{\boldsymbol{\beta}}_k^b - \widetilde{\boldsymbol{\beta}}_k)^{\mathsf{T}} \ . \tag{11}$$

Graphically, this corresponds to a data (or concentration) ellipsoid (Friendly *et al.*, 2011) of the bootstrap sample estimates, given by $\widetilde{\boldsymbol{\beta}}_k \oplus \widetilde{\boldsymbol{\Sigma}}_k^{1/2} \mathcal{S}$. Alternatively, robust versions of the empirical estimator in Eqn. (11) may easily be substituted, for example those based on the high-breakdown bound Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) methods developed by Rousseeuw and others (Rousseeuw and Leroy, 1987, Rousseeuw and Van Driessen, 1999).

For the present purposes, where the goal is simply to gauge the tradeoff of shrinkage vs. precision graphically, and where we are concerned with the *relative* sizes of the ellipsoids over choices of k (or df) rather than precise coverage, this naive bootstrap approach is usually sufficient. See Hall (1997, §4.2) for discussion of a wider range of alternatives for multivariate bootstrap regions.

Of course, the use of ellipsoids as a visual summary of the bootstrap estimates in this approach does not entirely remove the assumption of multivariate normality of the ridge estimates; it simply moves that assumption to that of the distribution of the bootstrap estimates $\widetilde{\boldsymbol{\beta}}_k^b$. If desired, this restriction can be removed by the use of non-parametric density estimation to construct smoothed approximations to the joint distribution of the $\widetilde{\boldsymbol{\beta}}_k^b$ following methods described, for example, by Hall (1987). In practical implementation, this idea is limited largely to 2D representations.

2.3 Details: Views

For a given data set, we thus have a set of K ellipsoids, $\mathcal{E}(\widehat{\boldsymbol{\beta}}_{k_j}^{\star}, \widehat{\boldsymbol{\Sigma}}_{k_j}^{\star})$, $j = 1, 2, \ldots, K$, each of dimension p. These can be displayed in a variety of ways. For example, in static displays, bivariate views of the 2D projections of these ellipsoids can be shown for given pairs of predictors superposing the ellipses for all values of k in each plot (as was done in Figure 2).

Alternatively, all superposed pairwise 2D views can be shown in a scatterplot matrix format (as in Figure 4). With suitable 3D software (e.g., the `rgl` package for R, Adler and Murdoch (2011)) similar overlaid plots for sets of three predictors can be obtained.

Finally, modern dynamic and interactive software provides other possibilities, including animating any of the above mentioned display formats over K, or providing interactive choices of views and/or the shrinkage or tuning factor via sliders or other software controls. I believe that such interactive implementations, which couple computation and display with interactive control are potentially quite useful. However, the present article stands as a proof-of-concept using static displays, leaving dynamic and interactive versions for future development.

2.4 Details: Reduced-rank views

As described above, the ellipsoids $\mathcal{E}(\widehat{\boldsymbol{\beta}}^{\star}_{k_j}, \widehat{\boldsymbol{\Sigma}}^{\star}_{k_j})$ may be viewed in the space of the predictor variables ($\boldsymbol{\beta}$ space) in several ways, but all of these are methods for showing p-dimensional effects in 2D (or 3D) views that can be seen on a screen or on paper. Unfortunately, these methods begin to suffer the curse of dimensionality as p grows large. For example the $p = 8$ plot (Figure 8) discussed in Section 3.2 approaches the limits of graphic resolution for the scatterplot matrix format.

As in other multivariate visualization problems, informative low-rank projections provide one antidote to the curse of dimensionality. In particular, Section 1.1 shows that the SVD transformation from the the column space of $\boldsymbol{X}$ to the orthonormal column space of $\boldsymbol{U}$, simplifies both computation and interpretation. Similarly, the same transformation can be applied to the ellipsoids $\mathcal{E}(\widehat{\boldsymbol{\beta}}^{\star}_{k_j}, \widehat{\boldsymbol{\Sigma}}^{\star}_{k_j})$, yielding a rotated, p-dimensional space, whose 2D projections in reduced-rank space can be particularly useful. The trick is to identify such informative 2D projections and be able to interpret them in terms of the original data.

Specifically, under a linear transformation by a conformable matrix $\boldsymbol{L}$, the image of the general ellipsoid $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\boldsymbol{L}(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \mathcal{E}(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^{\mathsf{T}}) = \boldsymbol{L}\boldsymbol{\mu} \oplus (\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^{\mathsf{T}})^{1/2} = \boldsymbol{L}\boldsymbol{\mu} \oplus \boldsymbol{L}\boldsymbol{\Sigma}^{1/2} \ . \tag{12}$$

Thus, taking $\boldsymbol{L} = \boldsymbol{V}$ gives views of the covariance ellipsoids $\boldsymbol{V}\widehat{\boldsymbol{\beta}}^{\star}_{k_j} \oplus \boldsymbol{V}\Sigma^{\star}_{k_j}{}^{1/2}$ of the ridge estimates in the space of the principal components of $\boldsymbol{X}$. Note that these ellipsoids will necessarily have their major/minor axes aligned with the coordinate axes in such plots, because the space of $\boldsymbol{V}$ is orthogonal and $\boldsymbol{V}$ is the matrix of eigenvectors of both $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ and $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}$.

2D plots with coordinate axes corresponding to the largest two singular values then show the effects of shrinkage in the subspace of maximum variance in $\boldsymbol{X}$. In general this is usually a good idea for multivariate visualization, but it turns out here to be uninformative or misleading, as illustrated below (Section 3.2). Instead, analogous plots in the subspace corresponding to the two *smallest* singular values give a view of ridge regression in the space where shrinkage is greatest, in a way similar to the collinearity biplot proposed by Friendly and Kwan (2009). For ease of interpretation, these plots may be supplemented with variable vectors defined by the rows of $\boldsymbol{V}$ showing the relations of the predictors to the reduced-rank space, as in a biplot(Gabriel, 1971). Our use of the term "biplot" here only connotes that such plots can show data summaries in one space (transformed $\boldsymbol{\beta}$ space) while simultaneously showing the projections of variable vectors into this space. Figure 3 shows an example of such a plot, whose interpretation is discussed in Section 3.1 below.

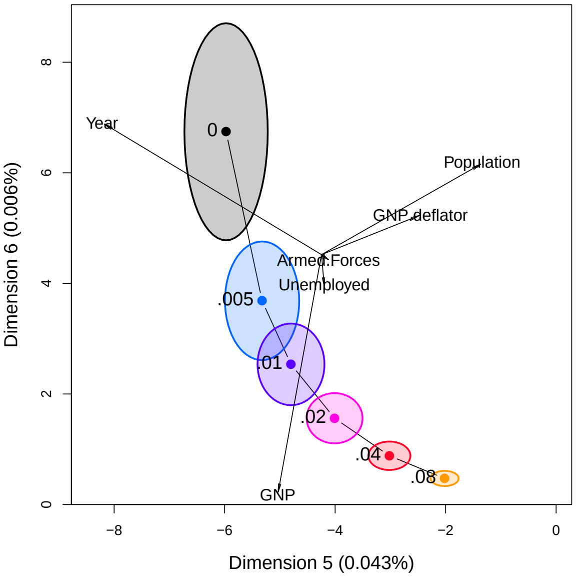

Figure 3: Ridge trace plots for the coefficients of predictors in Longley's data shown in the orthogonal space of the *smallest* two principal component vectors of $\boldsymbol{X}$, which contribute most to shrinkage. The variable vectors, positioned at an arbitrary origin and scaled to fill the available space, show the contributions of each variable to these dimensions. The plot uses an an aspect ratio of 1.0 to allow correct interpretation of lengths and angles.

2.5 Measuring precision and shrinkage

The multivariate extension of the ridge trace plot described above has another benefit in that it suggests simple ways to calculate summary measures of shrinkage and precision and thus provide other visualizations of the tradeoff, albeit with less detail.

From the theory above, shrinkage ("bias") can be measured by the length of the coefficient vector, $||\boldsymbol{\beta}_k|| = (\boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{\beta}_k)^{1/2}$. A normed version, $||\boldsymbol{\beta}_k|| / \max_k ||\boldsymbol{\beta}_k||$ will then give a relative measure with a maximum of 1 for $k = 0$.

As illustrated in our generalized ridge trace plots, we equate variance (inverse precision) with the "size" of the covariance ellipsoid of $\widehat{\boldsymbol{\Sigma}}_k$, and this can be quantified in several ways in terms of its eigenvalues $\boldsymbol{\lambda}_k = \boldsymbol{d}^2 / (\boldsymbol{d}^2 + k)$, a $p \times 1$ vector:

1. $\Pi_i \lambda_{k,i} = |\widehat{\Sigma}_k|$ measures the volume of ellipsoids, and corresponds conceptually to Wilks' Λ criterion in MANOVA. We prefer a linearized version, $\log(|\widehat{\boldsymbol{\Sigma}}_k|)$ or $|\widehat{\boldsymbol{\Sigma}}_k|^{1/p} = \sqrt[p]{\prod \lambda_i}$, the geometric mean.
2. $\sum_i \lambda_{k,i} = \text{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}})$ measures the average size over p dimensions and corresponds conceptually to the Pillai and Hotelling-Lawley trace criteria.
3. $\lambda_{k,1} = \max \boldsymbol{\lambda}_k$ corresponds to Roy's maximum root criterion.

Thus, a simple line plot of one of these measures of $\widehat{\Sigma}_k$ (or $\widehat{\Sigma}_k^{-1}$) versus $||\boldsymbol{\beta}_k|| / \max_k ||\boldsymbol{\beta}_k||$ will show directly the tradeoff of variance (or precision) against shrinkage, summarized across all predictors. See the example in Figure 6.

3 Examples

Two well-known, real-data examples of these methods are described below. For brevity, we discuss only the details of the data, statistical analysis and interpretation that are relevant to our graphical methods.

3.1 Longley data

Figure 2 uses the classic Longley (1967) data to illustrate bivariate ridge trace plots. The data consist of an economic time series ($n = 16$) observed yearly from 1947 to 1962, with the number of people Employed as the response and the following predictors: GNP, Unemployed, Armed.Forces, Population, Year, and GNP.deflator (using 1954 as 100). These data are often used as an example of extreme collinearity.

For each value of k, the plot shows the estimate $\widehat{\boldsymbol{\beta}}$, together with the covariance ellipse. For the sake of this example, we assume that GNP is a primary predictor of Employment, and we wish to know how other predictors modify the regression estimates and their variance when ridge regression is used.

For these data, it can be seen that even small values of k have substantial impact on the estimates $\widehat{\boldsymbol{\beta}}$. What is perhaps more dramatic (and unseen in univariate trace plots) is the impact on the size of the confidence ellipse. Moreover, shrinkage in variance is generally in a similar direction to the shrinkage in the coefficients. These effects provide a visual interpretation of Eqn. (4) as seen in bivariate views: The matrix $\boldsymbol{G}_k$ shrinks the covariance matrix of the OLS estimates in a similar way to the shrinkage of the estimates themselves by $\boldsymbol{G}_k$ in Eqn. (3).

Several other features, which cannot be seen in univariate ridge trace plots, are apparent in these plots. Most obvious is the fact that, for each pair of predictors, shrinkage of the coefficients follows a particular path through parameter (β) space. It can be shown (Friendly *et al.*, 2011) that the shrinkage path has a simple geometric interpretation as the *locus of osculation*[1] between two families of concentric ellipsoids: the elliptical contours of the covariance ellipsoid of the RSS function for OLS, and the spherical contours of the constraint term $k\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}$ in Eqn. (2). This path is the set of points where the normals to the two ellipsoids are parallel, and in the general case is given by a bi-quadratic form which plots as a conic section.

In Figure 4 it can be seen that the shrinkage paths are sometimes monotone in both parameters, but not always, as for the coefficients of Population and GNP. This occurs here because small values of the ridge constant initially drive the coefficient of Population *away* from zero while that of GNP goes toward zero, while larger values of k drive the coefficient of population back toward zero. See Jensen and Ramirez (2008) for discusssion of this and other anomalies in ridge regression.

Second, it can be seen that the covariance between the estimated ridge coefficients changes systematically along the ridge trace path. In all cases shown in Figure 2, the covariance decreases in absolute value with increasing k, though this is not a necessary feature. However, the essential feature to be seen here is that all covariance ellipsoids become smaller with increasing k, reflecting reduced variance or increased precision.

[1] The locus of osculation is the path of points along which two sets of curves just make contact ("kiss" or osculate), as

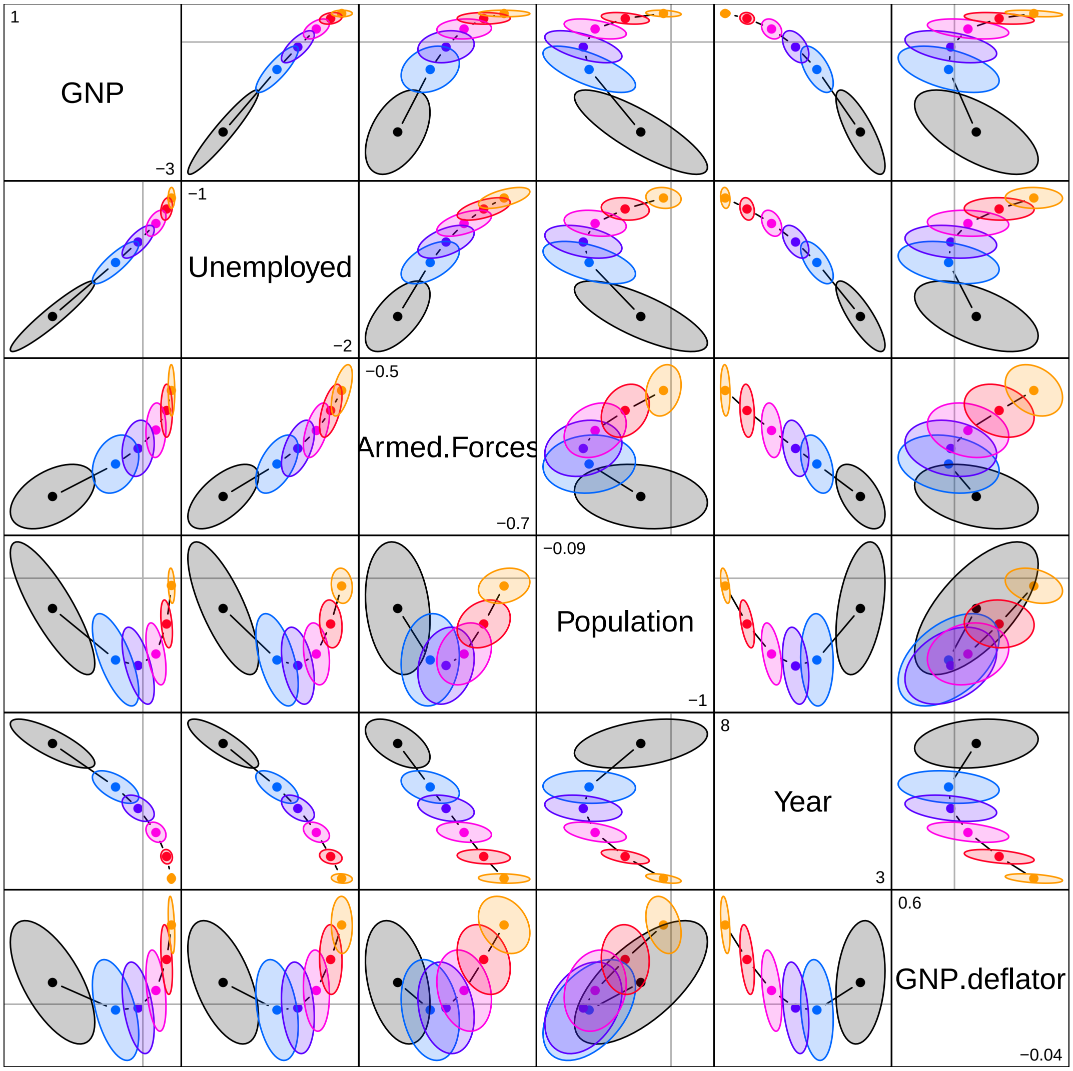

Figure 4: Scatterplot matrix of ridge trace plots for the coefficients of all predictors in Longley's data, with $k = 0, 0.005, 0.01, 0.02, 0.04, 0.08$. The same color coding as in Figure 2 is used here, with black showing the OLS estimates for $k = 0$.

Rather than showing selected bivariate plots as in Figure 2, some or all pairwise 2D views can be shown in a scatterplot matrix format as in Figure 4. Most details of our interpretation are similar to those above from Figure 2, except that it is now plainly seen that all predictors except for Population and GNP.deflator have a monotone pattern in their bivariate ridge paths. This nicely illustrates that it is the norm $||\boldsymbol{\beta}_k||$ which tends toward zero, not necessarily each coefficient, and that individual coefficients have different sensitivities over the range of k.

Moreover, in this format, it is easy to see the effect of the ridge constant on both bias and and variance jointly for a given variable, by scanning a given row or column in this pairwise display.

For comparison, Figure 5 shows the scatterplot matrix of all pairwise plots of the covariance el-

some parameter varies.

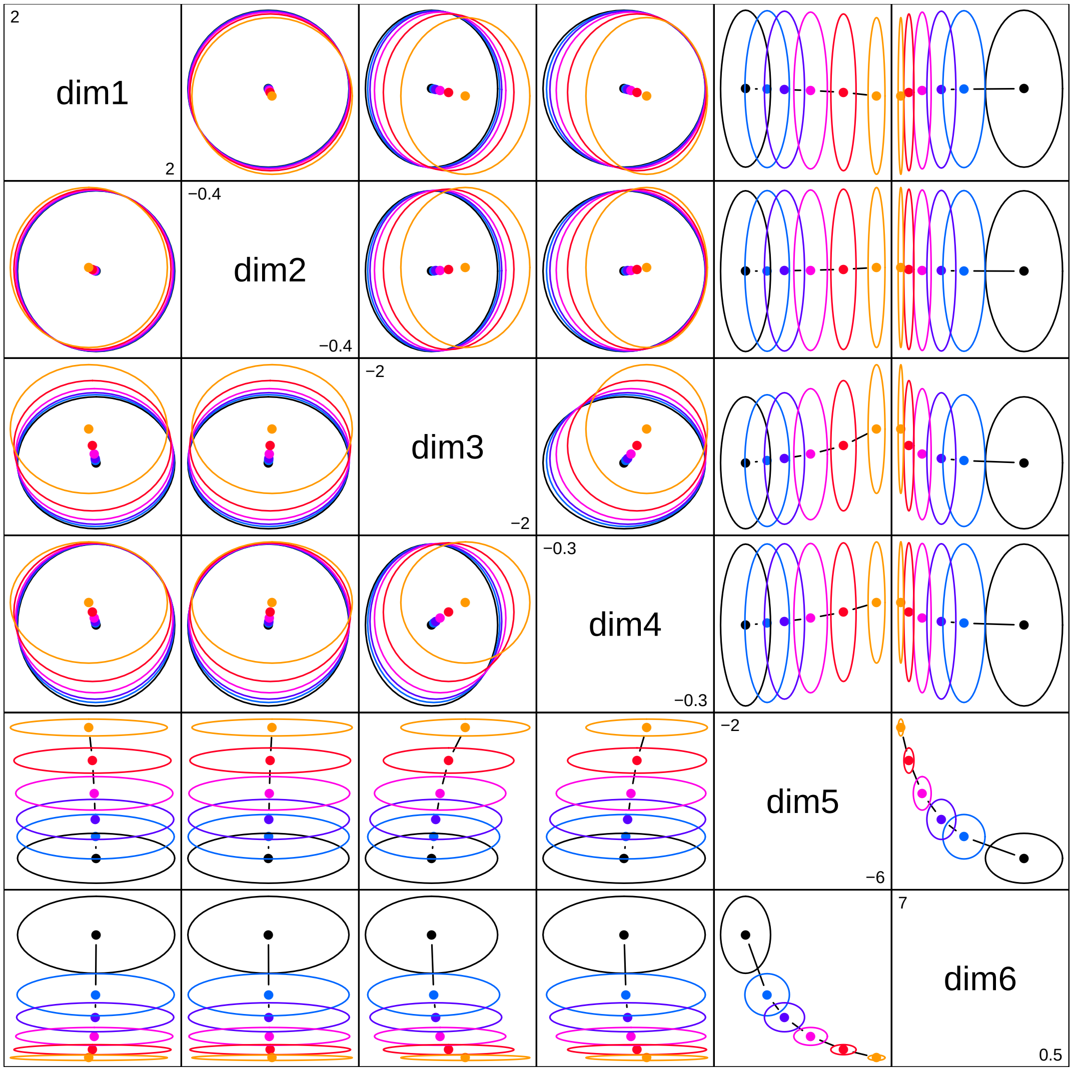

Figure 5: Scatterplot matrix of ridge trace plots for the coefficients of all predictors in Longley's data shown in the orthogonal space of the principal component vectors of $\boldsymbol{X}$. The plot makes clear that shrinkage occurs only in the space of the dimensions with the smallest eigenvalues (dim5 and dim6).

lipsoids transformed to the principal component space as in Eqn. (12). It is immediately clear that the shrinkage imposed by ridge regression takes place only in the last (smallest) two dimensions, a direct consequence of the shrinkage factors in Eqn. (6).

For interpretation of shrinkage in relation to the original variables, the most useful plot is the panel for the two smallest dimensions (dimensions 5 and 6 here), but annotated to show the variable vectors, as we showed in Figure 3. It is easily seen there that GNP, Year and Population contribute most to shrinkage along dimension 6 and that the greatest shrinkage occurs along this dimension. This is not surprising, given that the data are a time series over years. The reader may wish to compare this figure with the univariate ridge-trace plot in Figure 1. I hope you will agree that Figure 3 *does* provide direct visual evidence to decide about the tradeoff of shrinkage against precision.

Finally, for this example, Figure 6 illustrates two of the possible summary plots described in Sec-

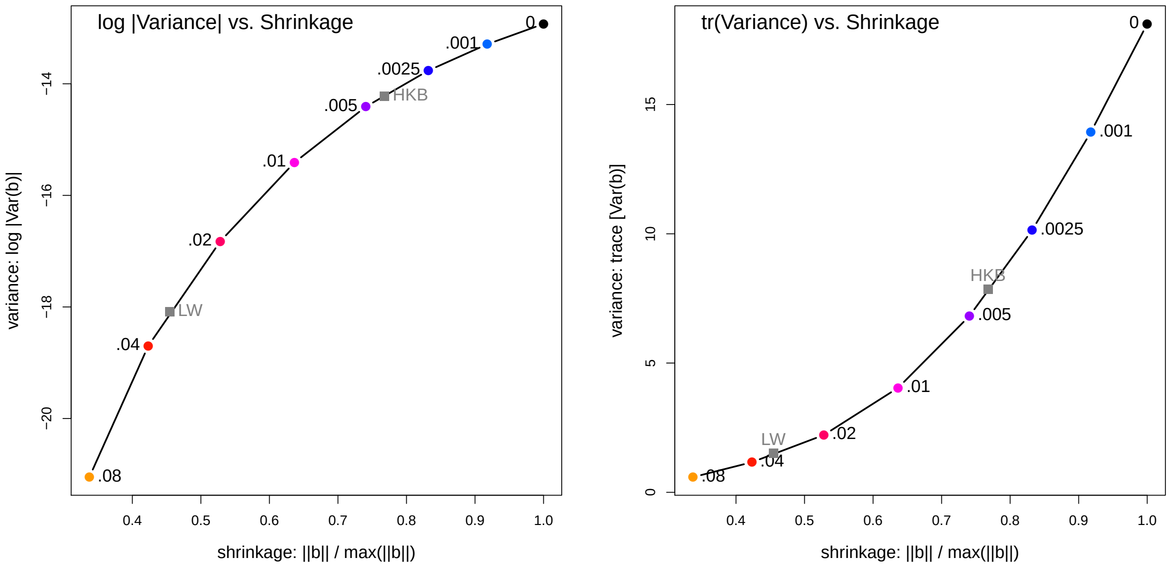

Figure 6: Variance versus shrinkage summary plots for Longley's data. Points along the curve show the "size" of the covariance matrix of parameters against the normed length of the coefficient vector, indexed by the shrinkage constant k. Left: using $\log(|\widehat{\boldsymbol{\Sigma}}_k|)$ as the measure of size; right: using $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}})$. Interpolated on the curves are the HKB and LW estimates of k as in Figure 1.

tion 2.5. In this plot, each ellipsoid has been summarized by the normed length of the coefficient vector, representing shrinkage relative to $||\boldsymbol{\beta}^{\mathrm{OLS}}|| = 1$ and either $\log(|\widehat{\boldsymbol{\Sigma}_k}|)$, or $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}})$. representing variance of the estimates. While lacking the rich detail of plots of the ellipsoids themselves, these plots shows how commonly used numerical criteria balance shrinkage against variance in this example. The Hoerl *et al.* (1975) (HKB) criterion favors relatively modest shrinkage, but achieves little in terms of reduced variance; the Lawless and Wang (1976) (LW) favors much greater shrinkage and gains considerably in reducing variance. However, the plots make clear that the tradeoff depends on the measure of "size" used to index variance.

3.2 Prostate cancer data

A second example uses data on prediction of the amount of the prostate specific antigen, used in diagnostic tests for prostate cancer, in a sample of 97 men about to undergo radical prostate surgery. The data come from a study by Stamey *et al.* (1989) and have been used extensively in the regression shrinkage and model selection literature following Tibshirani (1996), Hastie *et al.* (2001). The response is preoperative `lpsa`: log(prostate specific antigen), and the predictors are the following histological and morphometric measures: `lcavol`: log(cancer volume), `lweight`: log(prostate weight), `age`: patient age, `lbph`: log(benign prostatic hyperplasia), `svi`: seminal vesicle invasion, `lcp`: log(capsular penetration), `gleason`: Gleason grade of the prostate cancer, and `pgg45`: percentage of Gleason scores of 4 or 5.

Univariate trace plots for ridge regression applied to these data are shown in Figure 7. The left panel is the more traditional form, plotting coefficients versus k. The right panel parameterizes shrink-

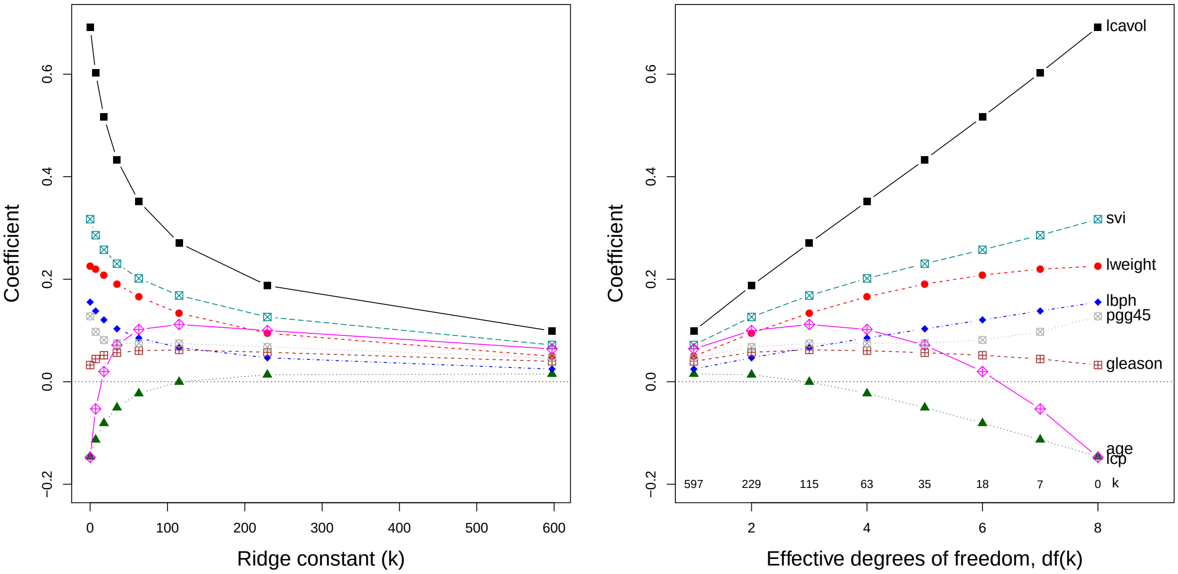

Figure 7: Univariate ridge trace plots for the coefficients of predictors predicting `lpsa` in the prostate cancer data. Left: traditional form, plotting coefficients versus k; right: the same coefficients versus $\mathrm{df}(k)$. The same graphic parameters (color, shape, line style) are used for the predictors in both plots, labeled in the right panel.

age in terms of $\mathrm{df}(k)$, and the nonlinear relation between the two tuning factors makes this plot more appealing and interpretable. But in either form they fail the interocular traumatic test: the purported message of a tradeoff between bias and variance is absent. Only the effects of bias are shown directly.

Figure 8 shows the bivariate, scatterplot matrix version of the ridge trace plot for all predictors in the linear model. Again, it is easy to see the joint effects on bias, variance and covariance of the estimated parameters by following a given variable in its row or column in the display.

Among other things, two variables stand out here, which are not apparent in the univariate views of Figure 7. The Gleason score (`gleason`) and percentage of Gleason scores of 4 or 5 (`pgg45`) both have non-monotonic bivariate traces with all other variables. As well, these are the variables with the largest relative variances, only reduced with large shrinkage. This is understandable, since the Gleason score is an ordered discrete variable with a range of 6–9, but only a few observations are outside 6–7. `pgg45` is based on the Gleason score and also exhibits large variance until the most extreme levels of shrinkage are approached.

As noted above in Section 2.4, these multivariate ridge trace plots can often be simplified by projection into subspaces of the principal components of the predictors. Figure 9 shows two such views for the current example. The left panel is a ridge-trace analog of a biplot (Gabriel, 1971), showing the covariance ellipsoids projected into the space of the first two principal components. The covariance ellipsoids are all aligned with the coordinate axes, allowing an easier interpretation of reduction in variance along these dimensions.

Overlaid on this plot are the variable vectors representing the columns of $\boldsymbol{V}$ in this space, positioned at an arbitrary origin. One interpretation is that shrinkage of the coefficients in Dimension 2 is most related to the variables `lbph`, `lweight`, and `age`, with other variables more related to

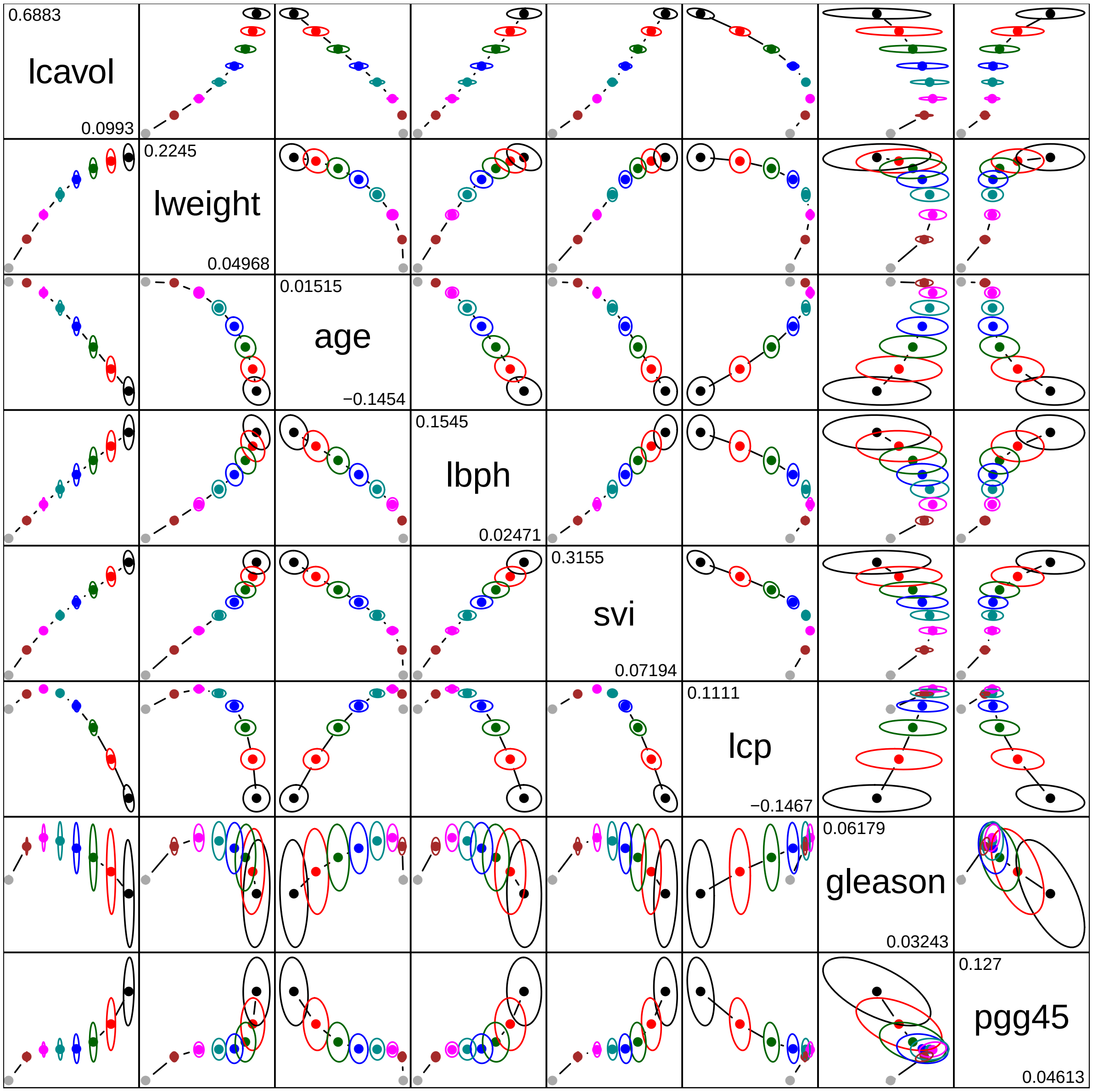

Figure 8: Scatterplot matrix of bivariate ridge trace plots for the coefficients of predictors in the prostate cancer data.

Dimension 1.

However, the right panel, showing the ridge-trace biplot in the space of the smallest two principal components is more relevant to shrinkage and selection problems, since it shows the relations of the variables to the directions of greatest shrinkage. It can be seen that the variables `ppg45`, `gleason` and `lcp` are most related to Dimension 8, while `svi` is also implicated in Dimension 7. The remaining variables, clustered at the origin of the variable vectors, have little relation to collinearity and shrinkage in this view.

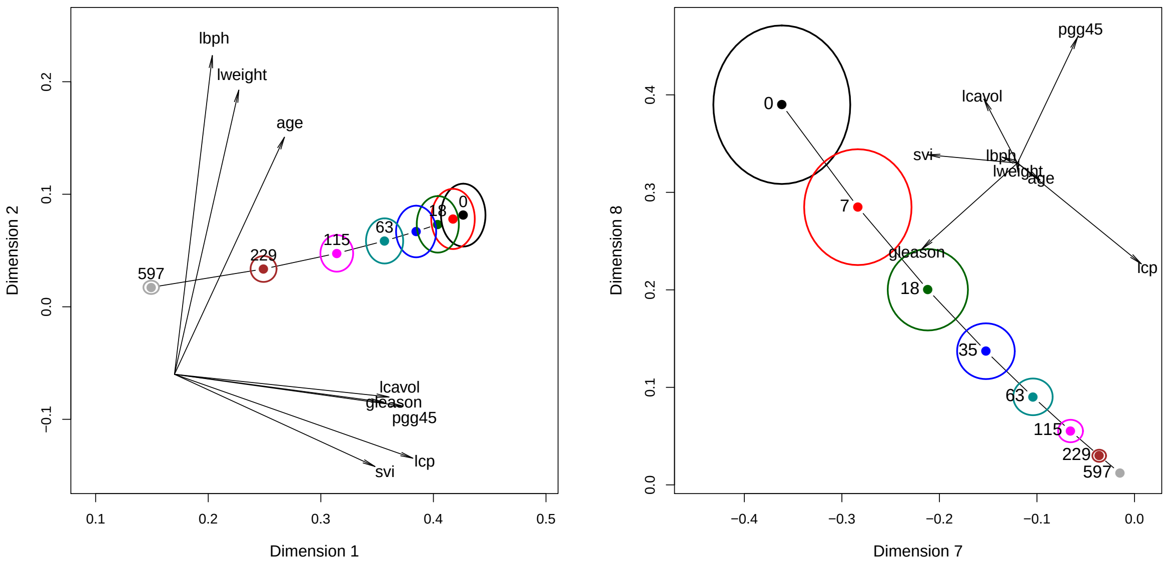

Figure 9: Reduced-rank ridge trace biplots for the prostate data. Left: dimensions 1 and 2; right: dimensions 7 and 8. Each plot has been scaled to an approximate aspect ratio of 1.0. The variable vectors show the relative projections of the original variables into this space from the corresponding columns of $\boldsymbol{V}$, positioned at a convenient arbitrary origin and scaled in length to fill the available space.

3.3 Extensions: Bootstrap methods

We return to the Longley (1967) data to illustrate the application of multivariate bootstrap methods for cases where analytic expressions for the covariance matrix $\boldsymbol{\Sigma}_k$ are unavailable and/or it is desired to avoid normality assumptions by replacing ellipsoids by non-parametric density estimates. As noted in Section 2.2, the latter is only computationally practical for 2D versions. Nevertheless, the examples below serve to provide some additional illumination for this process.

Figure 10 shows the results from an ordinary bootstrap resampling of the rows of the Longley data, generating $B = 800$ bootstrap samples and calculating the ridge regression estimates for the same values of k as in the previous examples ($k = \{0, 0.005, 0.01, 0.02, 0.04, 0.08\}$). With $p = 6$ predictors, and six values of k, each bootstrap sample gave a 6×6 matrix of ridge estimates, $\boldsymbol{\beta}_k, k \in K$. For simplicity in presentation, we only consider here the bivariate relations among two predictors: GNP and Unemployed, corresponding to the upper left panel in Figure 2.

The left panel of Figure 10 contains the data ellipses of the bootstrap estimates with the same radii ($c = 1/2 \approx \sqrt{2F_{.12}(2, 800)}$ as in the earlier figure. By and large, the size and orientation of the covariance ellipsoids from the bootstrap are consistent with those from the classical analytic estimates based on Eqn. (4). However, the individual bootstrap estimates for the OLS case ($k = 0$) are widely scattered, and give reason to worry about the adequacy of ellipsoids to capture the first and second moments of the multivariate bootstrap distribution for this problem.

The right panel of Figure 10 shows contours of the non-parametric 2D kernel density estimates for the bootstrap distribution, along the lines suggested by Hall (1987). It may be seen that the boot-

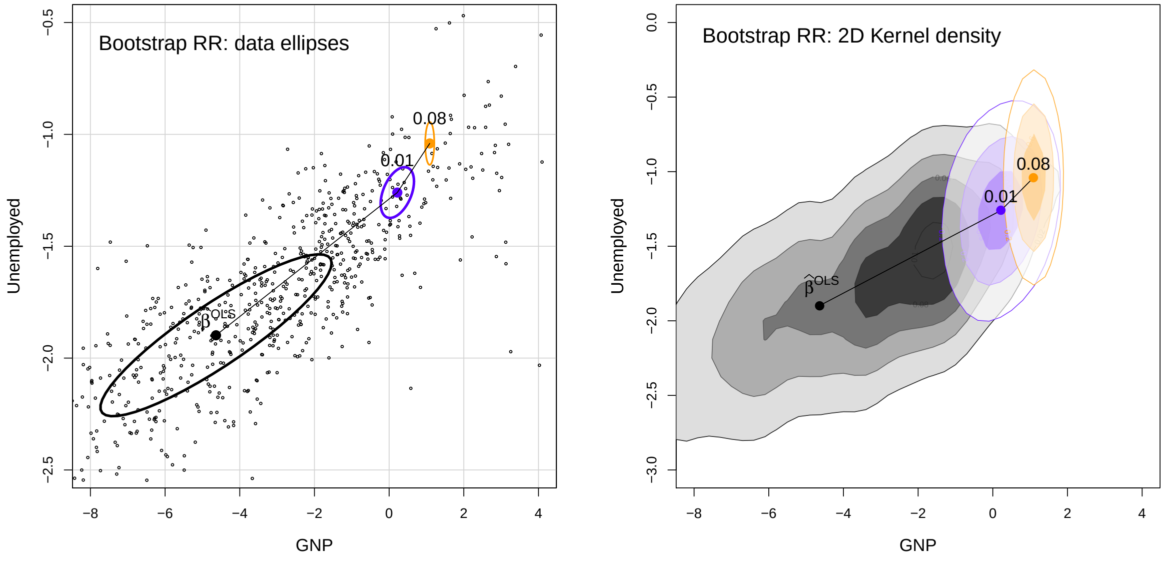

Figure 10: Results for $B = 800$ bootstrap samples of the ridge regression estimates for GNP and Unemployed in Longley's data. For simplicity, only the results corresponding to $k = 0, 0.01, 0.08$ are shown. Left: Individual bootstrap estimates are shown as points only for $k = 0$ (OLS), together with data ellipses of radius $c = 0.5$ computed using Eqn. (11) for all three shrinkage constants. Right: Contour plots of the 2D kernel density estimates computed with bandwidth 0.4 in each coordinate, using `bkde2D()` in the KernSmooth R package.

strap distribution of the OLS estimates differs markedly from elliptical and that the mode is rather far from the bootstrap estimate. However, in this and other examples we have tried, the contours of the shrunken estimates are more nearly elliptical. This suggests the conjecture that shrinkage, in addition to increasing precision, also improves the normal approximation on which these graphical methods rely.

4 Discussion

This paper makes two contributions to the literature on graphics for shrinkage methods, typified by ridge regression: (a) the development of a multivariate extension of the standard univariate ridge-trace plot using ellipsoids to show both bias and precision, and (b) the use of low-rank, 2D biplot projections to show informative views for higher-dimensional problems. With additional computational complexity, these ideas extend readily to other shrinkage methods for regression and under wider assumptions.

It arose as one example of the general idea that bivariate and multivariate views of data and statistical effects could in many cases be illuminated by the geometry of ellipses and ellipsoids under standard normal theory (Friendly *et al.*, 2011). Once this view was taken, it became clear why the widely used univariate ridge trace plot was a failure for its intended goal of showing the tradeoff of bias versus precision: the univariate plot of trace lines shows only the centers of the covariance ellipsoids ($\widehat{\boldsymbol{\beta}}^{\star}_k$), while information about precision is contained in the size and shape of $\Sigma^{\star}_k$.

Displaying both together in generalized ridge trace plots gives what should be the canonical view.

As we have shown, the mathematics and underlying p-dimensional geometry provide greater insight into the nature of shrinkage problems in regression. Moreover, as we have illustrated even the bivariate version of these plots show interesting features (non-monotonic bivariate trends, changes in the covariance as well as standard errors of estimated coefficients) not revealed in the univariate version. The reduced-rank, biplot views we have described provide one way to allow these graphic methods to extend easily to higher-dimensional problems.

5 Supplementary materials

All figures in this paper were constructed with R software (R Development Core Team, 2011). Functions implementing the graphical methods described here are included in the R package `genridge`, available on the CRAN site at `http://cran.r-project.org/package=genridge`. 2D and 3D plotting methods are provided, both in the space of the predictors and in transformed SVD/PCA space. Documentation examples for the principal functions in the package reproduce some of the figures shown here and include other data set examples as well. R code for all of the figures is included in the supplementary materials.

6 Acknowledgments

This work was supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada to Michael Friendly. I am grateful to John Fox for critical comments on an initial draft of this paper and to the associate editor and two reviewers for suggestions and comments that have helped to sharpen and extend the ideas presented.

References

Adler, D. and Murdoch, D. (2011). *rgl: 3D visualization device system (OpenGL)*. R package version 0.92.798.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.

Friendly, M. and Kwan, E. (2009). Where's Waldo: Visualizing collinearity diagnostics. *The American Statistician*, 63(1), 56–65.

Friendly, M., Monette, G., and Fox, J. (2011). Elliptical insights: Understanding statistical methods through elliptical geometry. Submitted, *Statistical Science*; available at `http://datavis.ca/papers/ellipses.pdf`.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics*, 58(3), 453–467.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76, 131–139.

Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74(3), 481–493.

Hall, P. (1997). *The Bootstrap and Edgeworth Expansion*. Berlin: Springer-Verlag, corr. 2nd printing. edn.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag Inc.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems (Corr: V12 p723). *Technometrics*, 12, 69–82.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4(2), 105–123.

Jensen, D. R. and Ramirez, D. E. (2008). Anomalies in the foundations of ridge regression. *International Statistical Review*, 76, 89–105.

Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, 5, 307–323.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62, 819–841.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), 93–100.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients. *Journal of Urology*, 141(5), 1076–1083.

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

Vinod, H. D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics*, 60(1), 121–131.