POINTS OF SIGNIFICANCE

Regularization

Constraining the magnitude of parameters of a model can control its complexity

Last month we examined the challenge of selecting a predictive model that generalizes well, and we discussed how a model's ability to generalize is related to its number of parameters and its complexity[1]. An appropriate level of complexity is needed to avoid both underfitting and overfitting. An underfitted model is usually a poor fit to the training data, and an overfitted model is a good fit to the training data but not to new data. This month we explore the topic of regularization, a method that controls a model's complexity by penalizing the magnitude of its parameters.

Regularization can be used with any type of predictive model. We will illustrate this method using multiple linear regression applied to the analysis of simulated biomarker data to predict disease severity on a continuous scale. For the *i*th patient, let y_i be the known disease severity and x_{ij} be the value of the *j*th biomarker. Multiple linear regression finds the parameter estimate $\hat{\beta}_j$ values that minimize the sum squared error $SSE = \Sigma_i(y_i - \hat{y}_i)^2$, where $\hat{y}_i = \Sigma_j \hat{\beta}_j x_{ij}$ is the *i*th patient's predicted disease severity. For simplicity, we exclude the intercept, $\hat{\beta}_0$, which is a constant offset. Recall that the hat on $\hat{\beta}_j$ indicates that the value is an estimate of the corresponding parameter β_j in the underlying model.

The complexity of a model is related to the number and magnitude of its parameters[1]. With too few parameters the model underfits, missing predictable systematic variation. With too many parameters it overfits, fitting training data noise with lower SSE but increasing variability in prediction of new data. Generally, as the number of parameters increases, so does the total magnitude of their estimated values (**Fig. 1a**). Thus, rather than limiting the number of parameters, we can control model complexity by constraining the magnitude of the parameters—a kind of limited budget for the model to spend on parameters. In fact, this can also reduce the number of variables in the model.

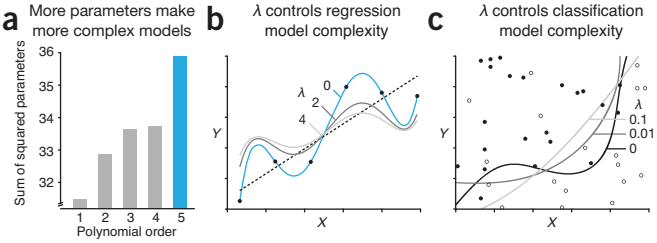

Figure 1 | Regularization controls model complexity by imposing a limit on the magnitude of its parameters. (**a**) Complexity of polynomial models of orders 1 to 5 fit to data points in **b** as measured by the sum of squared parameters. The highest order polynomial is the most complex (blue bar) and drastically overfits, fitting the data exactly (blue trace in **b**). (**b**) The effect of λ on ridge regression regularization of a fifth-order polynomial model fit to the six data points. Higher values of λ decrease the magnitude of parameters, lower model complexity and reduce overfitting. When λ = 0 the model is not regularized (blue trace). A linear fit is shown with a dashed line. (**c**) The effect of λ on the classification decision boundary of logistic regression applied to two classes (black and white). The regression fits a third-order polynomial to variables *X* and *Y*.

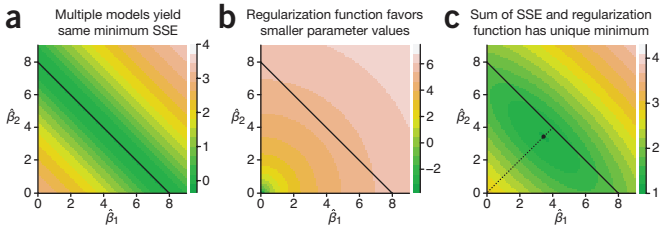

Figure 2 | Ridge regression (RR) can resolve cases where multiple models yield the same quality fit to the data. (**a**) The SSE of a multiple linear regression fit of the model $\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ for various values of $\hat{\beta}_1$ and $\hat{\beta}_2$. For each parameter pair, 1,000 uniformly distributed and perfectly correlated samples $x_1 = x_2$ were used with an underlying model of $y = 6x_1 + 2x_2$. The minimum SSE is achieved by all models that fall on the black line $\hat{\beta}_1 + \hat{\beta}_2 = 8$, shown in all panels. (**b**) The value of the RR regularizer function, $\lambda\Sigma_j\hat{\beta}_j^2$ ($\lambda = 9$), which favors smaller magnitudes for each parameter. (**c**) A unique model parameter solution (black point) found by minimizing the sum of the SSE and the regularizer function shown in **b**. As λ is increased, the solution moves along the dotted line toward the origin. Color ramp is logarithmic.

To see how this might work, let's consider a single patient and apply a three-biomarker model given by $\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. Suppose that, in reality, the underlying model—which we don't know—is $y = 5x_3$. The ideal estimate is $\hat{\beta} = (0,0,5)$. But what happens if, by chance, the values for the first two biomarkers in our training data are perfectly correlated; for example, $x_1 = 2x_2$? Now $\hat{\beta} = (50,-100,5)$ also gives the same fit as the ideal estimate because $50x_1 - 100x_2 = 0$. To put it more generally, as long as $\hat{\beta}_2 = -2\,\hat{\beta}_1$, the magnitude of $\hat{\beta}_1$ can be arbitrarily large. If x_1 and x_2 do not have perfect correlation in new data, models with nonzero values for $\hat{\beta}_1$ and $\hat{\beta}_2$ will perform worse than the ideal solution and, the larger their magnitudes, the worse the fit. By penalizing the magnitude of the parameters, we avoid models with large values of $\hat{\beta}_1$ and $\hat{\beta}_2$ and may even force them to their correct values of 0.

The classic approach to constraining model parameter magnitudes is ridge regression (RR), also known as Tikhonov regularization. It imposes a limit on the squared L^2-norm, which is the sum of squares of parameters, by limiting $M = \Sigma_j\hat{\beta}_j^2 \le T$. This is equivalent to minimizing $\Sigma_i(y_i - \hat{y}_i)^2 + \lambda\Sigma_j\hat{\beta}_j^2$. The second term is the regularizer function in which λ acts as a weight that balances minimizing SSE and limits the model complexity; the value of λ can be chosen. Achieving this balance is crucial; if we only care about model selection based on minimizing SSE on the training data, we typically wind up with an overfitted model. In general, the larger the value of λ, the lower the magnitude of parameters and thus model complexity (**Fig. 1b**,**c**). Note that even with large values of λ, parameter magnitudes are reduced but not set to zero. Thus, a complex model such as the fifth-order polynomial (**Fig. 1b**) is not reduced to a lower order polynomial. The relationship between *T* and λ is complex—depending on the data set and model, either *T* or λ may be more convenient to use. As we'll see below, a value of *T* directly corresponds to a boundary in the model parameter space that can be handily visualized.

One benefit of RR is that it can select a unique model in cases where multiple models yield an equally good fit, as measured by SSE (**Fig. 2a**, black line) by adding a regularizer function (**Fig. 2b**) to the SSE minimization to find a unique solution (**Fig. 2c**, black point). **Figure 2** is a simplification; in a realistic scenario the model

would have more parameters and correlation in the data would exist but would not be perfect. We show the regularization process for a fixed $\lambda = 9$ (**Fig. 2b**,**c**); the best value for λ would normally be chosen using a process like cross-validation that evaluates the model parameter solution using a validation data set[1].

An alternative regularization method is the least absolute shrinkage and selection operator (LASSO), which imposes a limit on the L^1-norm, $M = \Sigma_j|\hat{\beta}_j| \leq T$. This is equivalent to minimizing $\Sigma_i(y_i - \hat{y}_i)^2 + \lambda\Sigma_j|\hat{\beta}_j|$. Unlike RR, as λ increases (or T decreases), LASSO removes variables from the model by reducing the corresponding parameters to zero (**Fig. 3a**). In other words, LASSO can be used as a variable selection method, which removes biomarkers from our diagnostic test model.

To geometrically illustrate both RR and LASSO, we repeated the simulation from **Figure 2a** but without correlation in the variables (**Fig. 3b**). In the absence of correlation, the lowest SSE is at the parameter estimate $\hat{\beta} = (6, 2)$ which, in this example, happened to be the parameters used in the underlying model to generate the data. However, for our choice of T, this parameter estimate coordinate fell outside of the constraints of both regularization methods, and we instead had to choose a coordinate within the constraint space that had the lowest SSE. This coordinate corresponded to a model that yields a higher SSE than the minimum possible SSE but strikes a balance between model complexity and SSE. Recall that our goal here wasn't to minimize the SSE, which can easily be done by an overfitted model, but rather to find a model that would perform well with new data. In this example, we used T because it has a more direct geometrical interpretation; for example, it corresponds to the square of the radius of the RR boundary circle.

An interesting observation is that for some values of T, the LASSO solution may fall on one of the corners of the square boundary (**Fig. 3b**, $T = 3$). Since these corners sit on an axis where one of the parameters equals zero, they represent a solution in which the corresponding variable has been removed from the model. This is in contrast to RR, where because of the circular boundary, variables won't be removed except in the unlikely event that the minimum SSE already falls on an axis.

LASSO has several important weaknesses. First, it does not guarantee a unique solution (**Fig. 3c**). It is also not robust to colinearity in the data. For example, in a diagnostic test there may be several correlated biomarkers and, although we would typically want each of them to contribute to the model, LASSO will often select only one of them—this selection can vary even with minor changes in the data. Another potential weakness of LASSO manifests when the number of variables, P, is larger than the number of samples, N, a common occurrence in biology, especially with 'omic' data sets. LASSO cannot select more than N variables, which may be an advantage or a disadvantage depending on the data.

Given these weaknesses, an approach that blends RR and LASSO is often used for model selection. The elastic net method uses both regularization methods, simultaneously restricting both the L^1- and L^2-norms of the parameters. For linear regression, this is equivalent

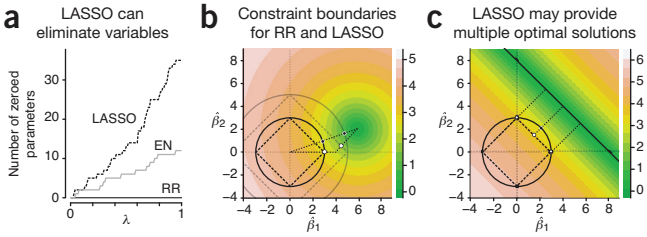

Figure 3 | LASSO and elastic net (EN) can remove variables from a model by forcing their parameters to zero. (**a**) The number of parameters forced to zero by each method as a function of λ. For each λ, 800 uniformly distributed samples were generated from a 100-variable underlying model with normally distributed parameters. For EN, an equal balance of the LASSO and RR penalties was used. (**b**) The same scenario as in **Figure 2b** but with independent and uniformly distributed x_1 and x_2. Constraint space and best solutions are shown for RR ($T = 9$, solid circle, black point) and LASSO ($T = 3$, dashed square, white point). Lighter circle and square correspond to RR with $T = 25$ and LASSO with $T = 5$. As regularization is relaxed and T is increased, the solutions for each method follow the dotted lines, and both approach the estimated parameters achieved without regularization: $\hat{\beta} = (6,2)$. (**c**) RR and LASSO constraint spaces shown as in **b** for the correlated data scenario in **Figure 2a**. RR yields a unique solution, while LASSO has multiple solutions that have the same minimum SSE. EN (not shown) also yields a unique solution, and its boundary is square with slightly bulging edges.

to minimizing $\Sigma_i(y_i - \hat{y}_i)^2 + \lambda_1\Sigma_j\hat{\beta}_j^2 + \lambda_2\Sigma_j|\hat{\beta}_j|$. This increases the number of models to be evaluated, as different combinations of λ_1 and λ_2 should be tried out during the cross-validation. If $\lambda_1 = 0$, we have LASSO; and if $\lambda_2 = 0$, we have RR. Elastic net is known to select more variables than LASSO (**Fig. 3a**), and it shrinks the nonzero model parameters like ridge regression[2].

Creating a robust predictor model requires careful control of the model complexity so that the model will generalize well to new data. This complexity can be controlled by reducing the number of parameters, but it is challenging to do so, as many combinations need to be evaluated. Regularization addresses this by creating a parameter budget used to prioritize variables in the model. Ridge regression provides unique solutions even when variables are multiply correlated, but it does not reduce the number of variables; while LASSO performs variable selection but may not provide a unique solution in every case. Elastic net offers the best of both worlds and can be used to create a simpler model that will likely perform better on new data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Jake Lever, Martin Krzywinski & Naomi Altman

1. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 703–704 (2016).
2. Zou, H. & Hastie, T. *J. R. Stat. Soc. B* **67**, 301–320 (2005).

Jake Lever is a PhD candidate at Canada's Michael Smith Genome Sciences Centre. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.