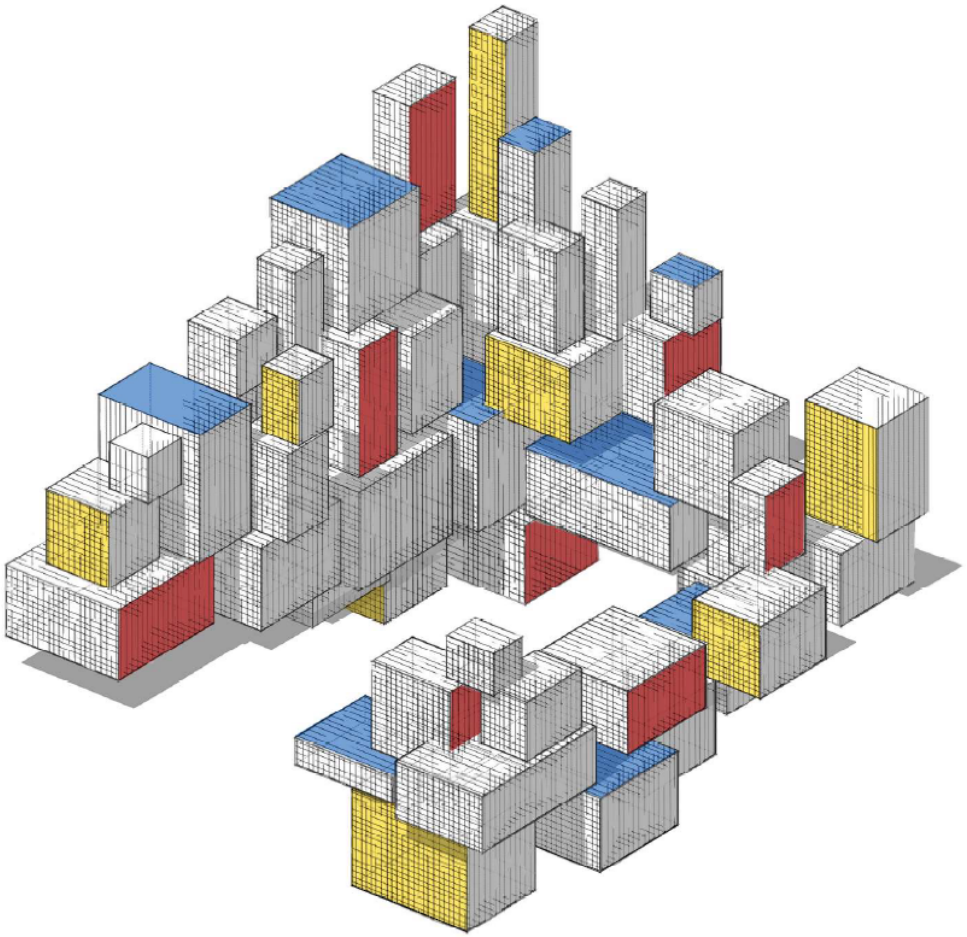
Regularization



Purpose

Kernel methods offer a broad generalization of linear models. Regularization provides a natural way to guard against overfitting. In this lecture we consider:

- a first view of kernel methods
- ridge regression
- lasso regression

Extending the Scope of Supervised Learning

We return to the supervised learning setting (regression) and expand its scope.

Given training data $\tau = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, we wish to find a prediction function (the learner) g_τ that minimizes the (squared-error) training loss

$$\ell_\tau(g) = \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i))^2$$

within a class of functions $\mathcal{G}$.

If $\mathcal{G}$ is the set of all possible functions then choosing any function g with the property that $g(\boldsymbol{x}_i) = y_i$ for all i will give zero training loss, but will likely have poor generalization performance (overfitting).

Recall that the best possible prediction function (over all g) for the squared-error risk $\mathbb{E}(Y - g(\boldsymbol{X}))^2$ is given by $g^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$.

Hilbert Space of Prediction Functions

The class $\mathcal{G}$ should be simple enough to permit theoretical understanding and analysis but, at the same time, rich enough to contain the optimal function g^* (or a function close to g^*).

This ideal can be realized by taking $\mathcal{G}$ to be a Hilbert space (i.e., a complete inner product space) of functions.

The set $\mathcal{G}$ of linear functions on $\mathbb{R}^p$ is a Hilbert space.

Namely, identify with each element $\boldsymbol{\beta} \in \mathbb{R}^p$ the linear function $g_{\boldsymbol{\beta}} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}$ and define the inner product on $\mathcal{G}$ as $\langle g_{\boldsymbol{\beta}}, g_{\boldsymbol{\gamma}} \rangle := \boldsymbol{\beta}^\top \boldsymbol{\gamma}$.

In this way, $\mathcal{G}$ behaves in exactly the same way as (is isomorphic to) the space $\mathbb{R}^p$ equipped with the Euclidean inner product (dot product). The latter is a Hilbert space, because it is complete with respect to the Euclidean norm.

Polynomial Regression Example

Let us now turn to our "running" polynomial regression example, where the feature vector $\boldsymbol{x} = [1, u, u^2, \ldots, u^{p-1}]^\top =: \boldsymbol{\phi}(u)$ is itself a vector-valued function of another feature u.

The space of functions $h_{\boldsymbol{\beta}} : u \mapsto \boldsymbol{\phi}(u)^\top \boldsymbol{\beta}$ is a Hilbert space, through the identification $h_{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}$.

This is true for *any* feature mapping $\boldsymbol{\phi} : u \mapsto [\phi_1(u), \ldots, \phi_p(u)]^\top$.

This can be further generalized to feature maps $u \mapsto \kappa_u$, where each κ_u is a real-valued function $v \mapsto \kappa_u(v)$ on the feature space.

Functions of the form $u \mapsto \sum_{i=1}^{\infty} \beta_i \kappa_{v_i}(u)$ live in a Hilbert space of functions called a reproducing kernel Hilbert space (RKHS).

Regularization

The aim of regularization is to improve the predictive performance of the best learner in some class of functions $\mathcal{G}$ by adding a penalty term to the training loss that penalizes learners that tend to overfit the data.

Let $\mathcal{G}$ be a Hilbert space of functions over which we search for the minimizer, g_τ, of the training loss $\ell_\tau(g)$.

Often, the Hilbert space $\mathcal{G}$ is rich enough so that we can find a learner g_τ within $\mathcal{G}$ such that the training loss is zero or close to zero, resulting in the risk of overfitting.

One way to avoid overfitting is to introduce a non-negative functional $J : \mathcal{G} \to \mathbb{R}_+$ which penalizes complex models (functions). In particular, we want to find functions $g \in \mathcal{G}$ such that $J(g) < c$ for some "regularization" constant $c > 0$.

Learning Problem

Thus we can formulate the quintessential supervised learning problem as:

$$\min \{\ell_\tau(g) \ : \ g \in \mathcal{G}, J(g) < c\}, \tag{1}$$

the solution (argmin) of which is our learner. When this optimization problem is convex, it can be solved by first obtaining the Lagrangian dual function

$$\mathcal{L}^*(\lambda) := \min_{g \in \mathcal{G}} \{\ell_\tau(g) + \lambda(J(g) - c)\},$$

and then maximizing $\mathcal{L}^*(\lambda)$ with respect to $\lambda \geqslant 0$.

In order to introduce the overall ideas of kernel methods and regularization, we will proceed by exploring (1) in the special case of ridge regression, with the following running example.

Ridge Regression

Ridge regression is simply linear regression with a squared-norm penalty functional (also called a regularization function, or regularizer).

We have a training set $\tau = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, with $\boldsymbol{x}_i \in \mathbb{R}^p$ and we use a squared-norm penalty with regularization parameter $\gamma > 0$.

Then, the problem is to solve

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i))^2 + \gamma \|g\|^2, \tag{2}$$

where $\mathcal{G}$ is the Hilbert space of linear functions on $\mathbb{R}^p$.

Ridge Regression

We can identify each $g \in \mathcal{G}$ with a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\|g\|^2 = \langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle = \|\boldsymbol{\beta}\|^2$.

The above functional optimization problem is thus equivalent to the parametric optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^2 + \gamma \, \|\boldsymbol{\beta}\|^2,$$

which further simplifies to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \| \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \gamma \, \|\boldsymbol{\beta}\|^2. \tag{3}$$

In other words, the solution to (2) is of the form $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ solves (3). Observe that as $\gamma \to \infty$, the regularization term becomes dominant and consequently the optimal g becomes identically zero.

Ridge Regression Solution

The optimization problem in (3) is convex, and by multiplying by the constant $n/2$ and setting the gradient equal to zero, we obtain

$$\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) + n\,\gamma\,\boldsymbol{\beta} = \mathbf{0}. \tag{4}$$

If $\gamma = 0$ these are simply the normal equations, albeit written in a slightly different form.

If the matrix $\mathbf{X}^\top\mathbf{X} + n\,\gamma\mathbf{I}_p$ is invertible (which is the case for any $\gamma > 0$), then the solution to these modified normal equations is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X} + n\,\gamma\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{y}.$$

Penalizing Specific Functions

When using regularization with respect to some Hilbert space $\mathcal{G}$, it is sometimes useful to decompose $\mathcal{G}$ into two orthogonal subspaces, $\mathcal{H}$ and $\mathcal{C}$ say, such that every $g \in \mathcal{G}$ can be uniquely written as $g = h + c$, with $h \in \mathcal{H}$, $c \in \mathcal{C}$, and $\langle h, c \rangle = 0$.

Such a $\mathcal{G}$ is said to be the direct sum of $\mathcal{C}$ and $\mathcal{H}$, and we write

$$\mathcal{G} = \mathcal{H} \oplus \mathcal{C}.$$

Decompositions of this form become useful when functions in $\mathcal{H}$ are penalized but functions in $\mathcal{C}$ are not.

We illustrate this decomposition with the ridge regression example where one of the features is a constant term, which we do not wish to penalize.

Ridge Regression (cont.)

Suppose one of the features in ridge regression is the constant 1, which we do not wish to penalize.

This ensures that when $\gamma \to \infty$, the optimal g becomes the "constant" model, $g(\boldsymbol{x}) = \beta_0$, rather than the "zero" model, $g(\boldsymbol{x}) = 0$.

Let us alter the notation slightly by considering the feature vectors to be of the form $\widetilde{\boldsymbol{x}} = [1, \boldsymbol{x}^\top]^\top$, where $\boldsymbol{x} = [x_1, \ldots, x_p]^\top$.

Let $\mathcal{G}$ be the space of linear functions of $\widetilde{\boldsymbol{x}}$; i.e., $g : \widetilde{\boldsymbol{x}} \mapsto \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$, so g is the sum of the constant function $c : \widetilde{\boldsymbol{x}} \mapsto \beta_0$ and $h : \widetilde{\boldsymbol{x}} \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}$.

Note that $\langle c, h \rangle = [\beta_0, \mathbf{0}^\top][0, \boldsymbol{\beta}^\top]^\top = 0$.

Ridge Regression (cont.)

As subspaces of $\mathcal{G}$, both C and $\mathcal{H}$ are again Hilbert spaces, and their inner products and norms follow directly from the inner product on $\mathcal{G}$.

For example, each function $h : \widetilde{\boldsymbol{x}} \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}$ in $\mathcal{H}$ has norm $\|h\|_{\mathcal{H}} = \|\boldsymbol{\beta}\|$, and the constant function $c : \widetilde{\boldsymbol{x}} \mapsto \beta_0$ in C has norm $|\beta_0|$.

The modification of the regularized optimization problem (2) where the constant term is not penalized can now be written as

$$\min_{g \in \mathcal{H} \oplus C} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\widetilde{\boldsymbol{x}}_i))^2 + \gamma \|g\|_{\mathcal{H}}^2,$$

which further simplifies to

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{n} \| \boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta} \|^2 + \gamma \|\boldsymbol{\beta}\|^2, \tag{5}$$

where $\mathbf{1}$ is the $n \times 1$ vector of 1s.

Ridge Regression (cont.)

Observe that, in this case, as $\gamma \to \infty$ the optimal g tends to the sample mean $\overline{y}$ of the $\{y_i\}$; that is, we obtain the "constant" model.

Again, this is a convex optimization problem, and the solution follows from

$$\mathbf{X}^\top(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) + n\,\gamma\,\boldsymbol{\beta} = \mathbf{0},$$

with

$$n\,\beta_0 = \mathbf{1}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}). \tag{6}$$

This results in solving for $\boldsymbol{\beta}$ from

$$(\mathbf{X}^\top\mathbf{X} - n^{-1}\mathbf{X}^\top\mathbf{1}\mathbf{1}^\top\mathbf{X} + n\,\gamma\,\mathbf{I}_p)\boldsymbol{\beta} = (\mathbf{X}^\top - n^{-1}\mathbf{X}^\top\mathbf{1}\mathbf{1}^\top)\boldsymbol{y}, \tag{7}$$

and determining β_0 from (6).

Gram Matrix

Suppose any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ can be written as a linear combination of the $\{\boldsymbol{x}_i\}$. So, $\boldsymbol{\beta} = \mathbf{X}^\top \boldsymbol{\alpha}$, for some $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\top \in \mathbb{R}^n$.

Then (7) reduces to

$$(\mathbf{X}\mathbf{X}^\top - n^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{X}\mathbf{X}^\top + n\,\gamma\,\mathbf{I}_n)\boldsymbol{\alpha} = (\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^\top)\boldsymbol{y}.$$

Assuming invertibility, we have the solution

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{X}\mathbf{X}^\top - n^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{X}\mathbf{X}^\top + n\,\gamma\,\mathbf{I}_n)^{-1}(\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^\top)\boldsymbol{y},$$

which depends on the training feature vectors $\{\boldsymbol{x}_i\}$ only through the $n \times n$ Gram matrix of inner products: $\mathbf{X}\mathbf{X}^\top = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$.

The solution for the constant term is $\widehat{\beta}_0 = n^{-1}\mathbf{1}^\top(\boldsymbol{y} - \mathbf{X}\mathbf{X}^\top\widehat{\boldsymbol{\alpha}})$. Hence, the learner is a linear combination of $\{\langle \boldsymbol{x}_i, \boldsymbol{x} \rangle\}$ plus a constant:

$$g_\tau(\widetilde{\boldsymbol{x}}) = \widehat{\beta}_0 + \boldsymbol{x}^\top\mathbf{X}^\top\widehat{\boldsymbol{\alpha}} = \widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i \,\langle \boldsymbol{x}_i, \boldsymbol{x} \rangle,$$

where the coefficients $\widehat{\beta}_0$ and $\widehat{\alpha}_i$ only depend on $\mathbf{X}\mathbf{X}^\top$.

Illustration of Ridge Regression Regularization

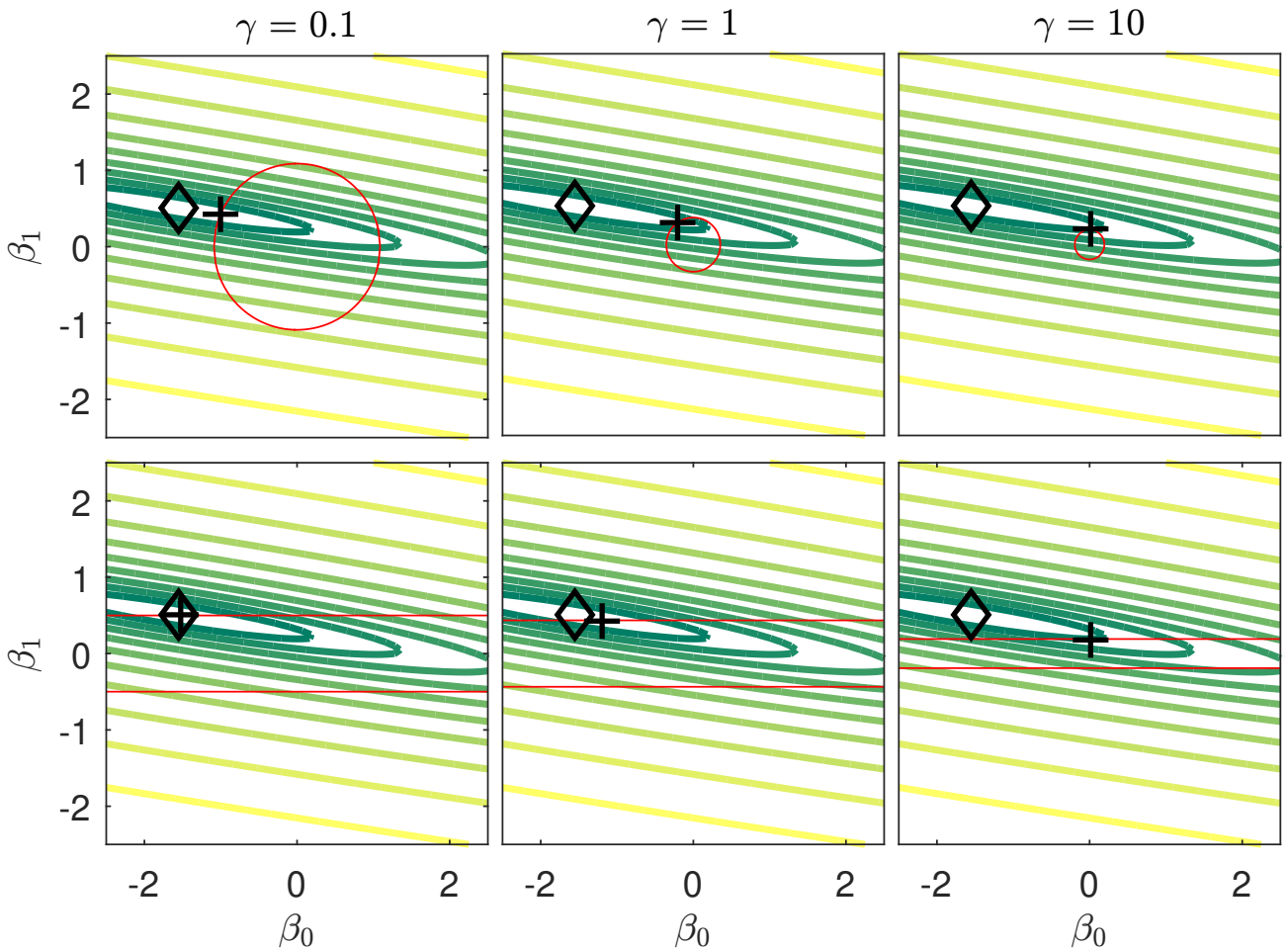
The following figure illustrates how the solutions of the ridge regression problems are qualitatively affected by the regularization parameter γ for a simple linear regression model.

The data was generated from the model $y_i = -1.5 + 0.5x_i + \varepsilon_i$, $i = 1, \ldots, 100$, where each x_i is drawn independently and uniformly from the interval $[0, 10]$ and each ε_i is drawn independently from the standard normal distribution.

In the figure:

- Contours are of the log squared-error loss
- Diamonds represent the minimizers
- Plusses show the minimizers of the regularized problems.

Each panel shows contours of the loss function (log scale) and the effect of the regularization parameter $\gamma \in \{0.1, 1, 10\}$.
Top row: both terms are penalized. Bottom row: only the non-constant term is penalized. Penalized (plus) and unpenalized (diamond) solutions are shown in each case.



Lasso Regression

Regularized ridge regression led to the minimization problem

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{n} \| \boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta} \|^2 + \gamma \|\boldsymbol{\beta}\|^2,$$

involving a squared 2-norm penalty $\|\boldsymbol{\beta}\|^2$.

Replacing the squared 2-norm with a 1-norm gives the lasso (least absolute shrinkage and selection operator). The lasso equivalent of this ridge regression problem is thus:

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{n} \| \boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta} \|^2 + \gamma \|\boldsymbol{\beta}\|_1, \tag{8}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$.

Lasso Regression

This is again a convex optimization problem.

Unlike ridge regression, the lasso generally does not have an explicit solution, and so numerical methods must be used to solve it.

Note that the problem (8) is of the form

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} \quad & f(\boldsymbol{x}) + g(\boldsymbol{z}) \\ \text{subject to} \quad & \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{z} = \boldsymbol{c}, \end{aligned}$$

with $\boldsymbol{x} := [\beta_0, \boldsymbol{\beta}^\top]^\top$, $\boldsymbol{z} := \boldsymbol{\beta}$, $\mathbf{A} := [\mathbf{0}_p, \mathbf{I}_p]$, $\mathbf{B} := -\mathbf{I}_p$, and $\boldsymbol{c} := \mathbf{0}_p$, and convex functions $f(\boldsymbol{x}) := \frac{1}{n} \| \boldsymbol{y} - [\mathbf{1}_n, \mathbf{X}] \, \boldsymbol{x} \|^2$ and $g(\boldsymbol{z}) := \gamma \|\boldsymbol{z}\|_1$.

There exist efficient algorithms for solving such problems, including the alternating direction method of multipliers (ADMM).

Illustration of the solutions of the lasso regression, taking the square roots of the previous regularization parameters.

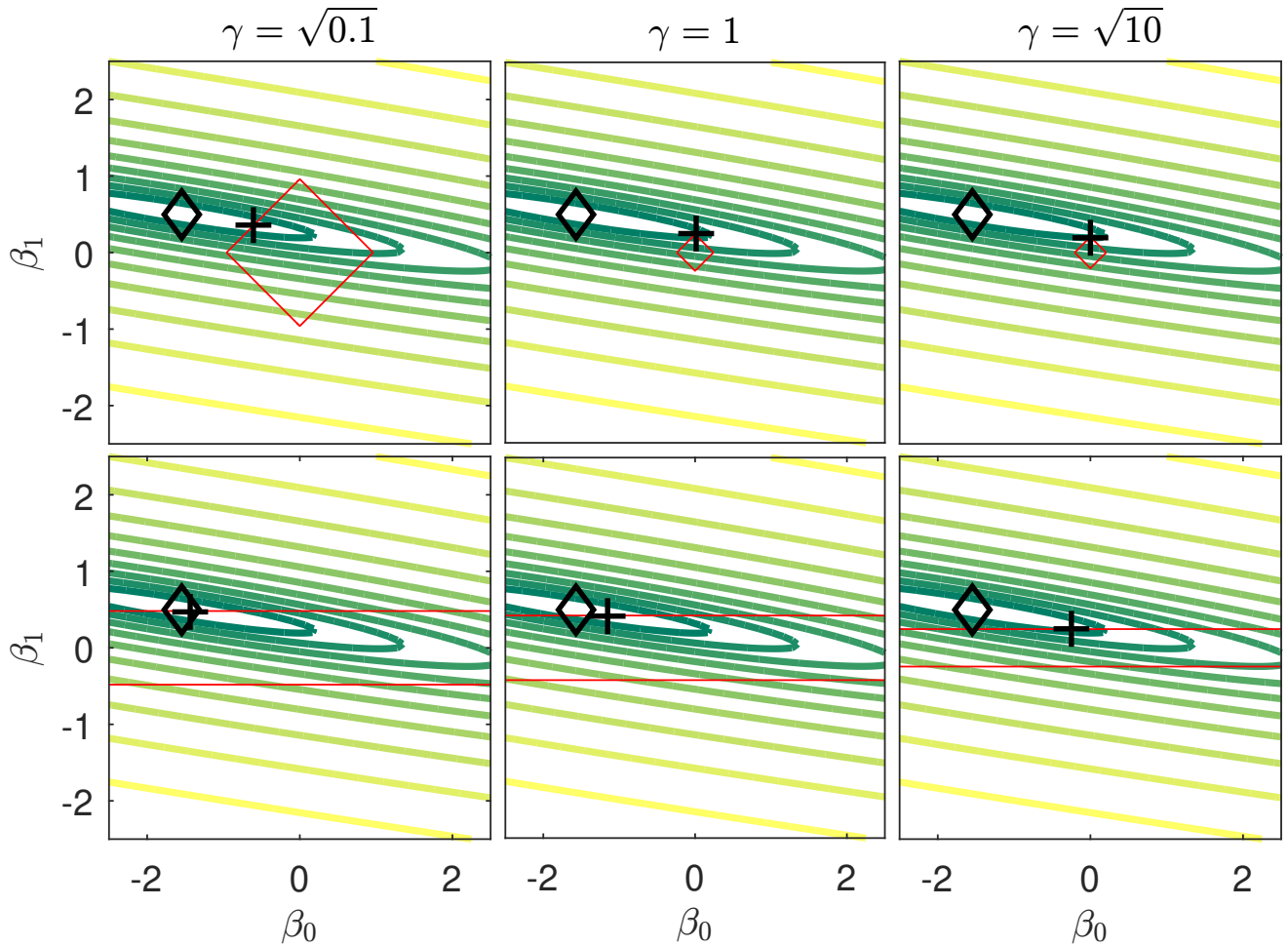


Figure: Lasso regression solutions.

Parsimonious Models

One advantage of using the lasso regularization is that the resulting optimal parameter vector often has several components that are exactly 0, leading to more parsimonious models.

For example, in the top middle and right panels of the previous figure, the optimal solution lies exactly at a corner point of the square $\{[\beta_0, \beta_1]^\top : |\beta_0| + |\beta_1| = |\beta_0^*| + |\beta_1^*|\}$; in this case $\beta_0^* = 0$.

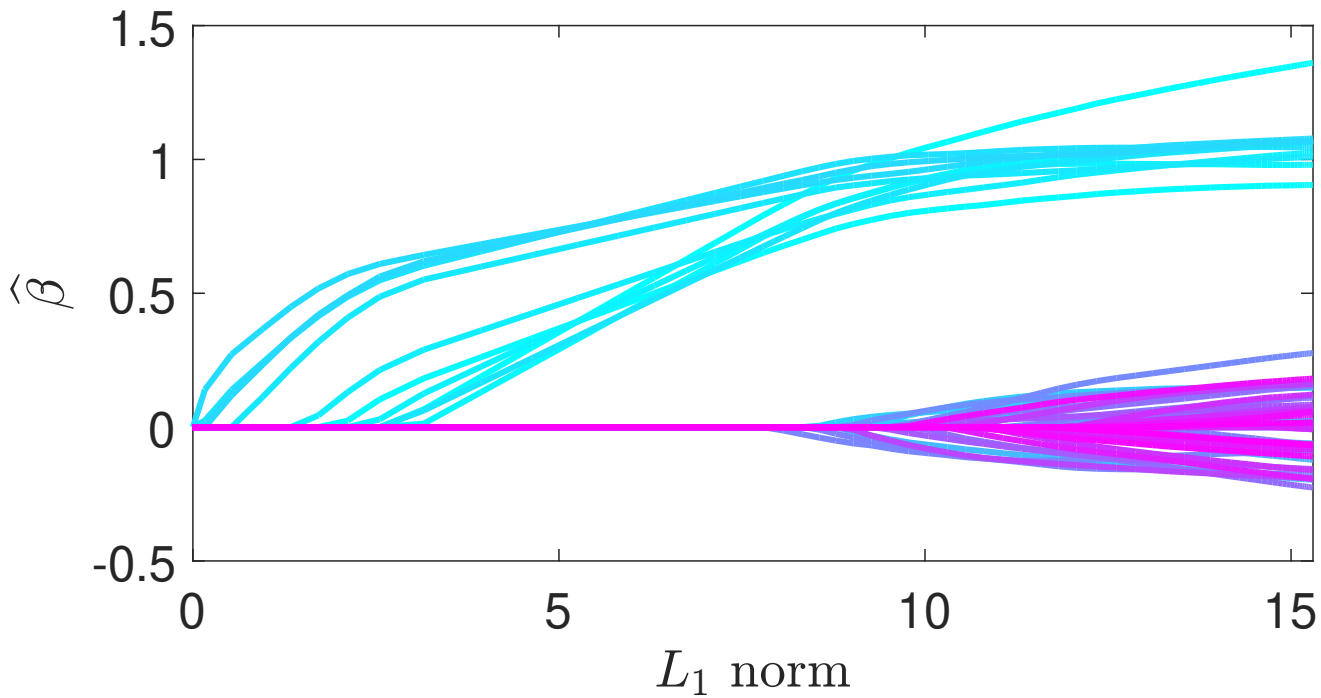
For statistical models with many parameters, the lasso can provide a methodology for model selection.

Namely, as the regularization parameter increases (or, equivalently, as the L_1 norm of the optimal solution decreases), the solution vector will have fewer and fewer non-zero parameters.

Regularization Paths

By plotting the values of the parameters against γ or the L_1 norm one obtains the so-called regularization paths (also called homotopy paths or coefficient profiles) for the variables.

Inspection of such paths may help assess which of the model parameters are relevant to explain the variability in the observed responses $\{y_i\}$.



Regularization Path

The following model was used:

$$Y_i = \sum_{j=1}^{60} \beta_j \, x_{ij} + \varepsilon_i, \quad i = 1, \ldots, 150,$$

where $\beta_j = 1$ for $j = 1, \ldots, 10$ and $\beta_j = 0$ for $j = 11, \ldots, 60$. The error terms $\{\varepsilon_i\} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Also $\{x_{ij}\} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

By the time the L_1 norm reaches around 4, all 10 variables for which $\beta_j = 1$ have been correctly identified and the remaining 50 parameters are estimated as exactly 0.

Only after the L_1 norm reaches around 8, will these "spurious" parameters be estimated to be non-zero.

The regularization parameter γ varied from 10^{-4} to 10.