Limitations of the Kernel trick

- Consider a regression with 2 predictors x_1, x_2. Model is $\hat{y} = X\hat{\beta}$ where $X = (1, x_1, x_2)$
- If we transform to $h(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$ then model becomes $\hat{y} = H\theta$ where θ is of length 6; we might estimate θ adaptively (eg via all subsets or lasso). We might leave out functions involving (eg) x_2.
- Polynomial kernel approach:

$$\begin{aligned} K(x, x') &= (1 + \langle x, x' \rangle)^2 = (1 + x_1 x_1' + x_2 x_2')^2 \\ &= (1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2') \end{aligned}$$

  If $h(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)$ then $K(x, x') = \langle h(x), h(x') \rangle$

- We can do ridge regression this way:

$$\begin{aligned} H\hat{\beta} &= H(H^T H + \lambda I_6)^{-1} H^T y \\ &= (HH^T + \lambda I_n)^{-1} HH^T y = (K(x, x') + \lambda I_n)^{-1} K(x, x') y \quad (1) \end{aligned}$$

- But this gives only linear shrinkage of the coefficients, can't adaptively leave out predictors. *Kernel trick does not work for adaptive methods like all-subsets, lasso etc.*