Causal Regularization

Dominik Janzing
Amazon Research Tübingen, Germany

Abstract

While regularization usually aims at avoiding to overfit finite data, I describe a scenario where regularization is even recommended in the infinite sample regime because it helps making predictive models more *causal*. To this end, I consider a d dimensional predictor linearly influencing a target variable. I describe a model of confounding that impacts the model parameters in exactly the same way as finite sample effects do. For this reason, keeping the penalizing term in Ridge and Lasso regression in the *infinite sample limit* tempers the impact of confounding equally well as it works against overfitting. Further, I prove a 'causal generalization bound' stating that any non-linear causal function 'generalizes' well from observational to interventional distribution whenever it is taken from a not too rich class, subject to my particular model of confounding.

Recall regularization for standard linear prediction tasks

Given a d-dimensional predictor variable $\mathbf{X}$ and real-valued target Y related by

$$Y = \mathbf{X}\mathbf{a} + E \quad \text{with } E \perp\!\!\!\perp \mathbf{X},$$

and $\mathbf{a} \in \mathbb{R}^d$. Ordinary least squares regression yields

$$\hat{\mathbf{a}} := \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}} = \mathbf{a} + \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}E}}.$$

where $\widehat{\Sigma_{\mathbf{X}E}} \neq 0$ due to finite sample effects.

Ridge and Lasso regression (Bayesian view)

For data matrices $\hat{\mathbf{X}}$ and $\hat{Y}$, Ridge and Lasso regression yield:

$$\hat{\mathbf{a}}_\lambda^{\text{ridge}} := \text{argmin}_{\mathbf{a}'}\{\lambda\|\mathbf{a}'\|_2^2 + \|\hat{Y} - \hat{\mathbf{X}}\mathbf{a}'\|^2\} \quad (1)$$
$$\hat{\mathbf{a}}_\lambda^{\text{lasso}} := \text{argmin}_{\mathbf{a}'}\{\lambda\|\mathbf{a}'\|_1 + \|\hat{Y} - \hat{\mathbf{X}}\mathbf{a}'\|^2\}, \quad (2)$$

where λ is a regularization parameter [1]. This can be justified via the priors

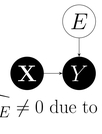
$$p_{\text{ridge}}(\mathbf{a}) \sim \exp\left(-\frac{1}{2\tau^2}\|\mathbf{a}\|^2\right) \quad (3)$$
$$p_{\text{lasso}}(\mathbf{a}) \sim \exp\left(-\frac{1}{2\tau^2}\|\mathbf{a}\|_1\right). \quad (4)$$

If we assume that the noise variable E is Gaussian with standard deviation σ_E we obtain

$$p(y|\mathbf{x},\mathbf{a}) \sim \exp\left(-\frac{1}{2\sigma_E^2}\|y - \mathbf{x}\mathbf{a}\|^2\right),$$

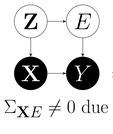
and then the posterior likelihood $p(\mathbf{a}|\hat{\mathbf{X}},\hat{Y})$ is maximized by the estimators (1) and (2).

Analogy between confounding and overfitting



$= \mathbf{X}\mathbf{a} + E$

$\widehat{\Sigma_{\mathbf{X}E}} \neq 0$ due to finite sample effects



$= \mathbf{X}\mathbf{a} + E$

$\Sigma_{\mathbf{X}E} \neq 0$ due to confounding

Both models induce the same distribution of covariance vectors:

$$\widehat{\Sigma_{\mathbf{X}E}} \sim \mathcal{N}(0, \widehat{\Sigma_{\mathbf{XX}}}\sigma_E^2/n) \qquad \text{while} \qquad \Sigma_{\mathbf{X}E} \sim \mathcal{N}(0, \gamma\Sigma_{\mathbf{XX}}),$$

for some γ if we choose a simple model of multivariate confounding with mixing from multiple independent sources [2] with random mixing vector $\mathbf{c} \sim \mathcal{N}(0, \sigma_c^2)$:

$\mathcal{N}(0, I) \sim$ 

$= \mathbf{Z}\mathbf{c}$

$\mathbf{Z}M =$

$= \mathbf{X}\mathbf{a} + E$

Regularization helps against both

For such a confounding model, Ridge and Lasso maximize posterior likelihood of $\mathbf{a}$ with non-zero λ in the population limit. But how to choose λ? Cross-validation is pointless.

Estimate confounding strength first

Define confounding strength [3]:

$$\beta := \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\hat{\mathbf{a}} - \mathbf{a}\|^2 + \|\mathbf{a}\|^2} \in [0,1]$$

Idea in [2]: the term $\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}E}$ tends to concentrate in the low eigenvalue subspace of $\Sigma_{\mathbf{XX}}$, while we assume $\mathbf{a}$ to be chosen from an isotropic prior.
The more $\hat{\mathbf{a}} = \mathbf{a} + \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}E}$ is concentrated in low eigenvalue subspace, the larger is β

Algorithm ConCorr

1: **Input:** i.i.d. samples from $P(\mathbf{X}, Y)$.
2: Rescale X_j to variance 1 if desired.
3: Compute the empirical covariance matrices $\widehat{\Sigma_{\mathbf{XX}}}$ and $\widehat{\Sigma_{\mathbf{X}Y}}$
4: Compute the ordinary least squares regression vector $\hat{\mathbf{a}} := \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}}$
5: Compute an estimator $\hat{\beta}$ for the confounding strength β via the algorithm in [2] from $\widehat{\Sigma_{\mathbf{XX}}}$ and $\hat{\mathbf{a}}$ and estimate the squared length of $\mathbf{a}$ via

$$\|\mathbf{a}\|^2 \approx (1 - \hat{\beta})\|\hat{\mathbf{a}}\|^2 \quad (5)$$

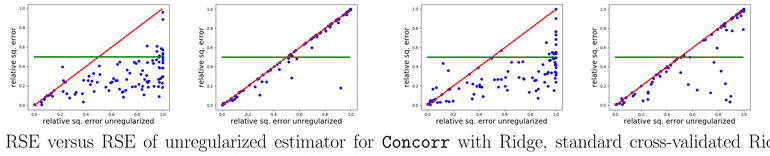
6: find λ such that the squared length of $\hat{\mathbf{a}}_\lambda^{\texttt{ridge/lasso}}$ coincides with the right hand side of (5)
7: Compute Ridge or Lasso regression model using this value of λ
8: **Output:** Regularized regression vectors $\hat{\mathbf{a}}_\lambda^{\texttt{ridge/lasso}}$

Simulation Results

Generate entries of $M, \mathbf{a}, \mathbf{c}$ from $\mathcal{N}(0,1), \mathcal{N}(0,\sigma_a^2), \mathcal{N}(0,\sigma_c^2)$ after σ_a, σ_c are uniformly drawn from $[0,1]$.

$$\mathbf{X} = M\mathbf{Z} \qquad Y = \mathbf{X}\mathbf{a} + \mathbf{Z}\mathbf{c} + E.$$

Define relative squared error $RSE := \frac{\|\hat{\mathbf{a}}_\lambda - \mathbf{a}\|^2}{\|\hat{\mathbf{a}}_\lambda - \mathbf{a}\|^2 + \|\mathbf{a}\|^2}$



RSE versus RSE of unregularized estimator for `Concorr` with Ridge, standard cross-validated Ridge, `ConCorr` with Lasso, standard cross-validaded Lasso for sample size 1000. `ConCorr` clearly outperforms cross-validation (for both Ridge and Lasso), which shows that cross-validation regularizes too weakly for causal modelling, as expected.

Real data

(1) Optical device with known ground truth [2]:

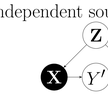
RSE with ConCorr (left: Ridge, right: Lasso) versus versus RSE of unregularized regression. Concorr helps in 3 out of 12 cases, and makes it worse in 0 cases.

(2) Taste of wine [2]:

11 ingredients influence the taste Y, alcohol is the main predictor. Dropping alcohol induces confounding. `ConCorr` with Ridge and Lasso yielded a relative error of 0.45 and 0.35, respectively, while [2] computed the confounding strength $\beta \approx 0.8$, which means that `ConCorr` significantly corrects for confounding.

Causal learning theory (non-linear functions)

Simple regression functions tend to better 'generalize' from *observational* to *interventional* distribution. Toy model of confounding with independent sources that shift Y:



$= Y' + \mathbf{Z}\mathbf{c}$

Confounding where $\mathbf{Z}$ influences Y in a linear additive way, while the influence on $\mathbf{X}$ is arbitrary. For some function f we define the **observational loss** measures its quality as predictive model:

$$\mathbf{E}[(Y - f(\mathbf{X}))^2] = \int [y - f(x)]^2 p(y, \mathbf{x}) d\mathbf{x} dy.$$

In contrast, the **interventional loss** measures the quality as causal model:

$$\mathbf{E}_{do(\mathbf{X})}[(Y - f(\mathbf{X}))^2] = \int [y - f(\mathbf{x})]^2 p(y|do(\mathbf{x})) d\mathbf{x} dy = \int [y - f(\mathbf{x})]^2 p(y|\mathbf{x},\mathbf{z}) p(\mathbf{z}) d\mathbf{z} d\mathbf{x} dy$$

Causal generalization bound

Introduce capacity measure for function class:

Definition 1 (correlation dimension) *Let $\mathcal{F}$ be some class of functions $f : \mathbb{R}^d \to \mathbb{R}$. Given the distribution $P_{\mathbf{X},\mathbf{Z}}$, the correlation dimension d_{corr} of $\mathcal{F}$ is the dimension of the span of*

$$\{\Sigma_{f(\mathbf{X})\mathbf{Z}} \,|\, f \in \mathcal{F}\}.$$

Theorem 1 (causal generalization bound) *Let $\mathbf{Z}$ be ℓ-dimensional with covariance matrix $\Sigma_{\mathbf{ZZ}} = \mathbf{I}$, influencing $\mathbf{X}$ in an arbitrary way Let the influence of $\mathbf{Z}$ on Y be given by a 'random linear combination' of $\mathbf{Z}$ with variance V according to the above additive model. Let $\mathcal{F}$ have correlation dimension d_{corr} and satisfy the bound $\|(f-g)(\mathbf{X})\|_{\mathcal{H}} \leq b$ for all $f \in \mathcal{F}$ (where $g(\mathbf{x}) := \mathbf{E}[Y'|\mathbf{x}]$). Then, for any $\beta > 1$,*

$$\mathbf{E}_{do(\mathbf{X})}[(Y - f(\mathbf{X}))^2] \leq \mathbf{E}[(Y - f(\mathbf{X}))^2] + b \cdot \sqrt{V \cdot \beta \cdot \frac{d_{\text{corr}} + 1}{\ell}},$$

holds uniformly for all $f \in \mathcal{F}$ with probability $e^{n(1-\beta+\ln\beta)/2}$.

Note: usual learning theory [4] relates *expected* loss to *empirical* loss plus a capacity term subject to exchangeability. Here we related observational to interventional loss subject to a confounder model with high symmetry properties.

Conclusions

The effect of multivariate confounders can be so similar to finite sample effects that the same techniques that avoid overfitting also temper the impact of confounding. Predicting target variables with functions from a 'small' function class increases the chances that the function also describes a causal relation.

The goal of this work is to stimulate a discussion on whether *causal* modelling requires stronger regularization than the one required by statistical predictability.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* Springer-Verlag, New York, NY, 2001.

[2] D. Janzing and B. Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

[3] D. Janzing and B. Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2017.

[4] V. Vapnik. *Statistical learning theory.* John Wileys & Sons, New York, 1998.