Tim B.

Formalizing the negative control method.

Let $X_1, \ldots, X_d$ be d predictors, $Y_1, \ldots, Y_p$ be p responses, and Z be a vector of confounders. Our objective is to test the hypothesis

$$X_i \perp\!\!\!\perp Y_j | Z$$

for (i, j) a preselected subset of the Cartesian product. We assume a Bayesian model for the X_is and Y_js. In particular, let $\theta_1^X, \ldots, \theta_d^X \sim G_X$ and $\theta_1^Y, \ldots, \theta_d^Y \sim G_Y$ be i.i.d. priors. We assume that the distributions $\mathcal{L}(X_i|Z)$ and $\mathcal{L}(Y_j|Z)$ have density functions $f_{\theta_i^X}(x_i|z)$ and $f_{\theta_j^Y}(y_j|z)$, respectively. If X_i and Y_j are independent given Z, then the density $f(x_i, y_j|z)$ is

$$f(x_i, y_j|z) = f_{\theta_i^X}(x_i|z) f_{\theta_j^Y}(y_j|z),$$

i.e., it is the product of the conditional distributions.

Connection to prediction.

Let $\{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$ be labeled training data. If our objective is prediction, we proceed as follows. First, select an algorithm A to apply to the data. In the case of prediction A is a predictive algorithm. Thus, our objective is for $\mathbb{E}[d(A(X_i, Z_i), Y_i)]$ to be small, i.e. to achieve small risk. (Here, d is a metric, for example $d(t_1, t_2) = (t_1 - t_2)^2$.) We select A so that this objective is satisfied on the training data.

Next, we have some holdout data on which we obtain an unbiased estimate of the algorithm A. Let $\{(X_{n+1}, Y_{n+1}, Z_{n+1}), \ldots (X_{n^*}, Y_{n^*}, Z_{n^*})\}$ be i.i.d. holdout data. Then

$$\sum_{i=n}^{n^*} d(A(X_i, Z_i), Y_i)$$

is our empirical estimate of the risk; we would expect risk to be similar on the unlabeled data.

We consider an analogy to conditional independence testing. For simplicity we consider a point null. Suppose that we want to test the hypothesis

$X \perp\!\!\!\perp Y | Z$. We have that $f(x, y, z) = f(z)f(x|z)f(y|z)$ under the null hypothesis (of course, we do not have access to these densities). Thus, suppose that we observe B datasets $\{D_1, \ldots, D_B\}$, where

$$D_i = \{(X_1^i, Y_1^i, Z_1), \ldots, (X_n^i, Y_n^i, Z_n)\}$$

is a dataset. (Thus, the datasets are i.i.d.). We could imagine generating these B synthetic datasets using a "symmetric CRT" style approach, i.e. we hold constant Z and we generate the Xs and Ys using the conditional density information $f(x|z)$ and $f(y|z)$. Note that we do *not* know $f(x|z)$ and $f(y|z)$, only that the data come i.i.d. from these distributions. (Put shortly, we observe B i.i.d. datasets from the null distribution). We select an algorithm A such that $A(D_i)$ is approximately uniform. Next, we generate B^* holdout datasets. On these B^* holdout datasets we run the algorithm A and assess

$$\left(\frac{1}{B^*} \sum_{i=1}^{B^*} p_i < \alpha \right) = \alpha$$

for all $\alpha \in (0, 1)$. (So, this is another kind of average.) Then, we run the algorithm on the unlabeled data.

We see that conditional independence testing (with negative controls) and prediction are analogous.

- **Training/calibration data**. data on which the ground truth is known that we use to tune/select our algorithm A.
  - Prediction: Labeled tuples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$.
  - CI testing: Datasets $\{D_i\}_{i=1}^B$ drawn from the null distribution.
- **Validation data**. Holdout data that we use to obtain an unbiased estimate of the performance of A.
  - Prediction: Additional labeled tubles $\{(X_i, Y_i, Z_i)\}_{i=1}^{n^*}$
  - CI testing: Additional datasets $\{D_i\}_{i=1}^{B^*}$ drawn from the null distribution.
- **Performance metric**. The metric that we use to gauge the performance of the algorithm.

  - Prediction: The risk. Let $m_{\mathrm{P}}(A) = \mathbb{E}\left[d(Y_i, A(X_i, Z_i))\right]$, where $d : \mathbb{R}^2 \to \mathbb{R}^{\geq 0}$ is some distance.
  - CI testing: The type-I error. Let $m_{\mathrm{CI}}(A, \alpha) = \mathbb{P}(A(D_i) \leq \alpha)$ for $\alpha \in (0, 1)$.

- **Objective**. Our target for the performance metric.
  - Prediction: We want $m_P(A)$ to be as small as possible.
  - CI testing: We want $|m_{\mathrm{CI}}(A, \alpha) - \alpha|$ to be as small as possible for all $\alpha \in (0, 1)$.

- **Empirical performance metric**. The sample analogue of the theoretical performance metric. We take an average over the data to compute the empirical performance metric for both prediction and CI testing. We use the validation data for an unbiased estimate.
  - Prediction:
$$\hat{m}_P(A) = \frac{1}{B^*} \sum_{i=1}^{n^*} d(Y_i, A(X_i, Z_i)).$$
  - CI testing:
$$\hat{m}_{\mathrm{CI}}(A, \alpha) = \frac{1}{B^*} \sum_{i=1}^{B^*} \mathbb{I}\left(A(D_i) \leq \alpha\right).$$
    Note that the corresponding visual diagnostic is a QQ-plot.

- **Unlabeled data**. Data on which the ground truth is unknown (i.e., the "real" data). After selecting/tuning the algorithm A and obtaining an unbiased estimate of $A's$ performance on the validation data, we apply A to the unlabeled data.

Ideally, we would have positive controls as well, which, under a similar Bayesian assumption, we could use to optimize power on the training data. In particular, suppose that the alternative distributions $f(x, y, z)$ come from some Bayesian model. Then, we might have the additional pieces.

- Datasets $\{S_i\}_{i=1}^{B'}$ drawn from the alternative distribution.

- An additional performance metric: power. Defined by

$$m'_{\mathrm{CI}}(A, \alpha) = \mathbb{P}(A(S_i) < \alpha).$$

- Empirical estimate of power:

$$\hat{m}'_{\mathrm{CI}}(A, \alpha) = \frac{1}{B'} \sum_{i=1}^{B'} \mathbb{I}\left(A(S_i) < \alpha\right).$$

In the training stage, we may want to optimize two criteria: empirical power and type-I error control. Positive controls may not always be available, in which case we simply can use the negative controls.

We therefore view CI testing as analogous to prediction. In both problems we pass through the following steps:

1. Select/tune algorithm on calibration data.
2. Estimate algorithm's generalization performance on validation data.
3. Apply algorithm to unlabeled data.

We can use other strategies from prediction as well. For example, if we have a small number of null datasets $\{D_i\}_{i=1}^{B^*}$, we might estimate the empirical type-I error (as a function of α) using cross-validation.

Note: the Bayesian assumption has the effect of converting a composite null into a point null, facilitating analysis.