Personal Notes from "Elements of Causal Inference"

by Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

Fall 2021

Contents

1 Pages 1-22

- Statistical learning VS causal inference:
  Statistical learning is about an inverse problem: estimating properties of the underlying distribution that can not be observed, based on the outcome of an operation (sampling) applied to it. Being an inverse problem, statistical learning is ill-posed, and therefore requires additional assumptions (e.g. class of functions). Causal learning implies multiple distributions, and it is ill-posed on two levels. It is still statistically ill-posed, but also ill-posed because even when the observational distribution is completely known, this might not be enough to determine the underlying causal model (structure learning / causal discovery / structure identifiability). This is because a causal structure contains more information than what is contained in the probability model alone, and statistical properties alone are not enough to determine causal structures.

- The existence of causal links can be inferred from statistical dependences:
  Reichenbach's common cause principle: A statistical dependence between X and Y indicates that they are caused by a confounder Z. This Z screens X and Y from each other, so that X and Y are conditionally independent given Z: $X \perp\!\!\!\perp Y|Z$. However:
  - the confounder Z can coincide with either X or Y
  - dependence might arise from selection bias
  - dependence might only be apparent (type I error in test, or i.i.d. assumption is violated)

- Two different causal models might induce the same observational distribution (i.e. be equal from a probabilistic point of view), but they can be distinguished from their intervention distributions:
  - A person is provided with class label y and produces the corresponding image x: X and Y are dependent, and Y causes X: $X = f(Y) + N_X$. Intervening on X does not change the corresponding value of Y. Intervening on Y does change X.
  - A person decides what to draw and produces both the label and the image: X and Y are dependent, as they are both function of the person's intention Z. Intervening on X does not change the corresponding value of Y. Intervening on Y does not affect X.

- Generic viewpoint assumption for causal inference: the observations is not due to chance, but it reflects properties of the objects. The viewpoint should not affect the property of the objects. (See Beuchet chair)

- **Principle of independent mechanisms**: The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. This means that the conditional distribution of each variable given its causes does not inform or influence the other conditional distributions. In the special case of 2 variables, this is called **independence of cause and mechanism (ICM)**. This principle implies that it is possible to performed a localized intervention ("tautology": when we intervene on X and change $p(x)$, nothing else happens). Also, p(effect|cause and p(cause) are autonomous / invariant mechanisms. Therefore, if a decomposition does not lead to autonomous mechanisms, the causal structure implied by that decomposition is not the correct one. This is because an effect obviously contains information about its cause, whereas the mechanism that generates the effect from its cause does not contain information about the mechanism that generates the cause.

2 Pages 23-39

- **Structural Causal Models (SCMs)**: SCMs are abstractions of underlying processes that take place in time and entail a joint distribution $P_{C,E}$ over causes and effects. Example:

$$C \to E \text{ (causal graph)}$$

$$C := N_C$$
$$E := f(C, N_E)$$

where f is a deterministic function and $N_C \perp\!\!\!\perp N_E$ (the noises are independent). Here, C is a direct cause of E.

Two SCMs with different canonical representation may induce the same interventional distribution but differ in the counterfactual statements.

- Interventions: partly change the data generation process, thus inducing a new distribution.
  - hard intervention: replace an assignment (i.e. replace an equation in the SCM)
  - soft intervention: keep the SCM while changing the noise distribution

An intervention on a cause C changes the distribution of its effect E, whereas an intervention on the effect E breaks the dependence between C and E, without changing the distribution of C in any way.

  - Example 3.2:

$$C \to E$$

$$C := N_C$$
$$E := 4C + N_E$$

with $N_C \perp\!\!\!\perp N_E$, $N_C, N_E \sim \mathcal{N}(0,1)$.

$\mathbb{E}[C] = 0$ and $\mathrm{Var}[C] = 1$.
$\mathbb{E}[E] = 0$ and $\mathrm{Var}[E] = 17$.

Intervention: set $C := 2$. Then $\mathrm{Var}[E] = 1$ (no contribution from C to the variance, since C is fixed), $\mathbb{E}[E] = 8$. The distribution of E changed.

Intervention: set $E := 2$. This has no impact on C, so $N_C \sim \mathcal{N}(0,1)$ as before and $P_C = \mathcal{N}(0,1)$. However, originally $P(C|E=2) \neq \mathcal{N}(0,1)$.

$P(C|E=2)$ can be computed with Bayes formula from Gaussian distribution:
$P(E|C) = \mathcal{N}(4C, 1)$
$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$
$P(C|E=2) = \frac{P(E=2|C)P(C)}{P(E=2)} = ...$

- Counterfactual: a modification of an SCM that changes all of its noise distributions. This preserves the physical mechanisms. Here the modification corresponds to conditioning the distribution, and therefore the structure of assignments is preserved. Counterfactuals can be used to falsify the causal model ("knowing the outcome of action x, what if instead...?"). Two SCMs with different canonical representations may induce the same interventional distributions, but they might still differ in the counterfactual statements (those depend on the noise).

3 Pages 43-62

- Structure identifiability requires additional assumptions:
  The causal structure in 2-variable settings can not be recovered from their joint distribution or from conditional independence test. In the bivariate case, for all joint distributions $P_{X,Y}$ over X, Y there always exists a SCM $X \to Y$, but also one $Y \to X$ ($\exists$ SCMs in both directions).

- **A-priori restriction of the model class** is a possible way out: since one of the functions describing the SCM in the "opposite" direction would be very complicated, we can restrict the model class (compare to regularization in ML). Models that hold under different background conditions (better generalization) are more likely to be causal.

- Special cases where identifiability is "guaranteed"/proved:

  - LiNGAMs:
    For linear models, if
    $Y = \alpha X + N_Y$, $N_Y \perp\!\!\!\perp X$
    there exists β s.t. $C = \beta E + N_C$ iff N_E, C are Gaussian.
    Since the existence of linear SCMs in both direction is "iff" noise and cause are jointly Gaussian, then Linear **Non-Gaussian** models are identifiable. In other words, even though a resonable fit is possible in both directions, the correct SCM can be found from the residuals (they should be independent of Y).
  - ANMs:
    $Y = f_y(X) + N_Y$, $N_Y \perp\!\!\!\perp X$
    A distribution $P_{X,Y}$ does not admit an ANM in both directions at the same time (generally).
  - Discrete ANMs: (similar to ANMs)
  - Post-nonlinear models $Y = g_y(f_y(X) + N_Y)$, $N_Y \perp\!\!\!\perp X$: a post-nonlinear model exists at most in one direction (again some exception)
  - Information-geometric (IGCI) models: identifiability from covariance / covariance matrix. The idea is that strong assumptions (deterministic relation in both direction, differentiability, strictly positive density, etc.) lead to "independence in one direction implies dependence in the other", which allows to identify the correct SCM.

4 Pages 62-79

- Additive Noise Models:

  Identifiabilty:$Y = f_Y(X) + N_Y$, $N_Y \perp\!\!\!\perp X$ only holds in one direction, not both.

  Causal discovery: From independence of residuals. If not conclusive, favor the highest p-value. Or by maximum-likelihood: compare $L_{X\to Y} = -\log \hat{\text{var}}[X] - \log \hat{\text{var}}[R_Y]$ and $L_{Y\to X} = -\log \hat{\text{var}}[Y] - \log \hat{\text{var}}[R_X]$. Which one is higher?

  [Comment: Independence of residuals should also work for LiNGaM. LinGaMs are skipped here because they are similar to ANMs (a special case of them)?]

- Information-Geometric Causal Inference:

  Assume a deterministic relation in both directions: $Y = f(X)$ and $X = f^{-1}(X)$.

  The noise is constant.

  f is differentiable, bijective, with differentiable inverse, strictly monotonic on $[0, 1]$.

  P_X has strictly positive continuous density.

  Independence condition: $\text{cov}[\log f', p_X] = 0 \implies C_{X\to Y} \leq C_{Y\to X}$ and $H(X) \leq H(Y)$ (differential Shannon entropy).

  Identifiability: $\text{cov}[\log f^{-1'}, p_Y] \geq 0$.

  Causal discovery: The variable with larger entropy is assumed to be the cause (or $X \to Y$ when $\hat{C}_{X\to Y} < \hat{C}_{Y\to X}$) .

  [Comment: Shannon entropy measures the uncertainty of a random process / the amount of information in the data. My interpretation of entropy in causality: the effect is more deterministic than the cause?]

- Trace method (for high-dimensional variables with linear relationship):

  $\mathbf{Y} = A\mathbf{X} + N_X$, $N_X \perp\!\!\!\perp \mathbf{X}$

  $\tau_e(A\Sigma_{XX}A^T) = \tau_d(\Sigma_{XX})\tau_e(AA^T)$ (renormalized traces): eigenvalues of Σ_{XX} are uncorrelated with the effect of A on eigenvectors.

  Identifiability: $\tau_d(A^{-1}\Sigma_{YY}A^{-T}) \leq \tau_d(\Sigma_{YY})\tau_d(A^{-1}A^{-T})$

  Causal discovery: Tracial dependency ratio: $r_{X\to Y} = \frac{\tau(A_Y \Sigma_{XX} {A_Y}^T)}{\tau(\Sigma_{XX})\tau({A_Y}^{-1}{A_Y}^{-T})}$ closer to 1 identifies the direction.

- Machine learning approach:

  Training data: each point is a dataset, each label is the causal direction.

  We can't use symmetric features!

  We can use entropy estimates of marginal distributions or entropy estimates for distribution of residuals (similar to ANM and LiNGaM).

- SSL in causal direction will not work: additional x's only tell about P_X, which does not give any information about $P_{Y|X}$ because of independent mechanisms. However, since P_{cause} and $P_{\text{effect|cause}}$ change independently, in the anticausal direction P_{effect} and $P_{\text{cause|effect}}$ are dependent.

5 Pages 81-96

- 4 elements of a causal model (Figure 6.2):
  - a corresponding complete DAG, where parents are called direct causes and children are called direct effect of their parents.
  - an observational distribution: sample idd from the noise (jointly independent noise variables), and use the structural assignments from the source nodes to the parents (Code 6.4). There must be a source, because the graph is DAG.
  - intervention distributions: replace one or several structural assignments
  - counterfactuals
- Structural minimality: remove redundancy by requiring that the functions depend on all their inputs. This leads to a unique minimal representation.
- Interventions change the distribution of the system. Intervention on variable X removes the causal influence of X's parents, and causes changes on X's children.
  - Atomic (deterministic) intervention: put a point mass on a real value: do(X=a).
  - Imperfect intervention: direct causes stay
- Example 6.10: $X_1 \to Y \to X_2$. X_2 is a better predictor for Y than X_1 is: anticausal direction and machine learning (see previous Chapters).
- Intervention is different from conditioning: intervention can change the data generation, conditioning filters data from observed underlying distribution (intervention on the effect can break the dependency with the causes).
- Total causal effect from X to Y: we randomize X, but we find that X and Y are dependent on each other (example: amount of drug, recovery). Alternative equivalent definition: we can find two different values of X such that the distribution of Y changes in different ways. It is not enough that there is a directed path between X and Y: if there is more than one path, two paths can cancel each other.

6 Pages 96-109

- Counterfactuals: incorporate the observed data into the distribution (i.e. update the noise distribution by conditioning on observation) and then analyze intervention.
  - Example 1:
    $X = N_X$,
    $Y = X^2 + N_Y$,
    $Z = 2Y + X + N_Z$.
    Observation: $(X, Y, Z) = (1, 2, 4)$
    Therefore: $(N_X, N_Y, N_Z) = (1, 1, -1)$
    Counterfactual: with intervention do$(X = 2)$, then $(Y, Z) = (5, 11)$
  - Probabilistically and interventionally equivalent but not counterfactually equivalent SCMs. Example 2:
    $X_1 = N_1$,
    $X_2 = N_2$,
    $X_3 = (1_{N_3>0}X_1 + 1_{N_3=0}X_2)1_{X_1!=X_2} + N_3 1_{X_1=X_2}$
    and
    $X_1 = N_1$,
    $X_2 = N_2$,
    $X_3 = (1_{N_3>0}X_1 + 1_{N_3=0}X_2)1_{X_1!=X_2} + (2 - N_3)1_{X_1=X_2}$
    have the same graphical causal model $X_1 \to X_3 \leftarrow X_2$,
    have the same observational distribution,
    Assume observation $(X_1, X_2, X_3) = (1, 0, 0)$. Then $(N_1, N_2.N_3) = (1, 0, N_3 \leq 0)$. With intervention $X_1 = 0$, $(X_1, X_2, X_3) = (0, 0, N_3)$ or $(X_1, X_2, X_3) = (0, 0, 2 - N_3)$.
- Markov properties:
  - a distribution entailed from an SCM is Markovian wrt the graph: d-separation by a set implies independence given the set that d-separates (from graph to distribution)
  - each node is independent of its non-descendants given its parents $\to$ if nodes are dependent, there must be a causal explanation, either one causes the other or there is a confounder that causes both (Reichenbach common cause principle). Graphically, dependence implies the existence of a non-blocked path.
  - Markov blanket of a node: parents, children, and parents of the children excluding itself
  - factorization: the joint probability is the product of conditional probabilities $p(node|parents)$ (but source nodes have no conditioning)
    [Comments: The factorization property connects to reinforcement learning: given the factorization, what would happen if ... (intervention on a state)? For example, what is maximum expected reward?]
  - Markov equivalence classes: equivalent DAGs have same skeleton and same immorality (v-structure)
- Faithfulness is the reverse implication of global Markov property: independence given a set of nodes implies d-separation by that set (from distribution to graph). It is not the same as causal minimality, and causal minimality is much weaker (see p.109): faithfulness implies causal minimality (but not the other way around)
- Causal minimality: a graph that is Markovian wrt to the graph, but not Markovian wrt to any subgraph (removing one edge corresponds to a new conditional independence that does not hold in the distribution)

- Recap:

  Markov property: d-separation implies independence given the set that d-separates

  Markov property implies Reichenbach common cause, because if we test and find dependence, there must be a $\leftarrow$ or $\rightarrow$ or a confounder, which means d-connected.

  So, assume we test and find dependence instead: with Markov property alone, this tells us there is no d-separation.

  But if we find independence, Markov property alone does not tell us anything. We need the other direction.

  The other direction is faithfulness: independence given a set implies d-separation by that set. This gives us an iif statement.

  So now, conditional independence test guarantees us d-separation. This means we have a way for causal discovery because:

  1) find all conditional independences
  2) draw the DAG that satisfy them all (all the d-separations)

  But still, there might be more than one graph because we are only guaranteed to find one representative in the equivalence class.

7 Pages 109-120

- Autonomy: causal relationships are autonomous under interventions, meaning that if we intervene on a variable, the other mechanisms are unchanged.
- Truncated factorization allows to compute statements about intervention distributions. Using autonomy property from above, the density after intervention can be computed with Markov property for all the variables where there has not been any intervention.
- Valid adjustment set: A set of nodes is a valid adjustment set for the pair (X, Y) if the adjustment formula (as in Kidney stones example) holds. Adjusting for Z allows to compute the average causal effect. If a set is a valid adjustment set, the conditionals remain the same also after intervening on X (invariant conditionals): $p^{do(X:=x)}(y|x,z) = p(y|y,z)$ and $p^{do(X:=x)}(z) = p(z)$.
- Valid adjustment sets for (X,Y) are parents of X, or backdoor, or "toward necessity".
- Kidney stones example:
  - First, we want to compare the probability of recovery (R=1) for T=A and for T=B independently from the size Z: $P^{do(T:=A)}(R=1)$ and $P^{do(T:=B)}(R=1)$. However, the observed data are "biased" on the size Z. The size Z has no parents.
    $P^{do(T:=A)}(R=1) = \sum_z P(R=1|T=A,Z=z)P(Z=z)$ (adjusting for the variable Z)
  - Then, $P^{do(T:=A)}(R=1) - P^{do(T:=B)}(R=1)$ is the average causal effect (ACE). This needs to be computed from do(), not from conditioning (conditioning is equivalent to empty adjustment set). If $p^{do(X:=x)}(y) \neq p(y|x)$, the causal effect is confounded. If the valid adjustment set is empty, the causal effect from X to Y is unconfounded.
- Or, intervention distributions can be computed from observational distribution and the graph, if they are identifiable. An intervention distribution is identifiable if there is a valid adjustment set.
- Front-door adjustment: If we do not observe a confounder, we can't apply the backdoor criterion, and there is no adjustment set. But still, we can use Markov property (why it is called "do-calculus" in 6.23?).

8 Pages 135-154

- Structure identifiabilty (recovering the graph from the joint distribution):

  Given a distribution that is Markovian wrt graph, then there exists a SCM with that graph. And given a complete DAG, there exists a corresponding SCM. And graphs are in Markov equivalence classes. So, more assumptions needed.

- Faithfulness: (see session 6) If a distribution is Markovian and faithful, then d-separation iff conditional independence. So in theory we could find all conditional independences and draw **a** DAG that satisfies them all (= satisfies all the d-separations). However, we will not able to distinguish between Markov equivalent graphs.

- Gaussian distributions: the causal effect can be summarized by a single number. However, if we only know the Markov equivalence class, we don't know that number.

- Additive Noise Models: additive noise with strictly positive densities, differentiable functions, causal minimality (not constant dependencies). If the distribution is induced by a linear Gaussian SCM, there is no guarantee of recovering the correct graph. Otherwise, we have identifiability. In particular, linear Gaussian with equal error variances, linear non-Gaussian, and non-linear Gaussian are identifiable.

- Learning causal relations from samples from different environments (interventional settings).

  Interventional equivalent classes of graphs: include intervention node with no parents, its children are the variables we intervene on. This increases the number of v-structures. Two graphs are intervention equivalent if they have same skeleton and same immoralities.

- Learning only partial causal structure (i.e. causal parents of some node Y): Y conditional on its parents is invariant under all interventions, if Y has not been intervened on.

- Independence based algorithms: test for conditional independences to find d-separation in the graph, assuming Markov condition and faithfulness.

  First, estimate the undirected skeleton. Two nodes are adjacent iff they cannot be d-separated (i.e. if they are always dependent). If two nodes are not adjacent, they are d-separated by the parents of one of them. Immoralities are easy to orient, once d-separation is given. Then, orient other edges to avoid cycles.

  However, statistical significance tests are asymmetric, the sample size is finite, and alpha is not really a parameter to tune.

- SAT satisfiability, Greedy Search, BIC, maybe dynamic programming, etc.

9 Pages 157-177

- Half-Sibling Regression: exploit the causal structure to reconstruct an unobserved signal Q that causes Y. Y has Q and N as parents. N is also parent of X. Then Y can be denoised by taking out all the part of Y that can be explained by X (by regressing Y on X). What is left is an estimate for Q. Here, X provides information on Q other than the one provided by Y because of causal faithfulness. Faithfulness says that independence given a set implies d-separation by that set. (And remember the iif with Markov property). But here Q and X are not d-separated by Y, because there is an unblocked path Q $\rightarrow$ Y $\leftarrow$ N $\rightarrow$ X. So Q is not independent of X given Y.

- Inverse Probability Weighting: construct an estimation by exploiting the fact that the density of a SCM and of one of its intervention distribution factorize in the same way besides for the term of the intervened variable.

- Episodic Reinforcement Learning: the estimator obtained as above can be used to estimate the performance of a different strategy.

- Domain Adaptation: a different domain corresponds to a different environment, which in Chapter 7 corresponded to an intervention on any node (not the target node Y). By invariant prediction property, Y conditional on its parents is invariant under all interventions, if Y has not been intervened on. However, there might be other sets than the set of parents that satisfy this. In covariate shift, there is always some set for which Y conditional on that set is invariant. If we don't know the set, the method of invariant causal prediction gives the intersection of all sets that satisfy invariance.

- Causal sufficiency: no hidden common cause is causing more than one variable
  VS
  Interventional sufficiency: there exists an SCM that induces observational and intervention distributions that coincide with what we observe. Basically, it corresponds to a set that is large/good enough to do causal inference. Sometimes, a set is interventionally sufficient but not causally sufficient: we can compute intervention distributions but there are hidden confounders. So, causal sufficiency is stronger.

- Two-stage least squares with instrumental variables: $Y = \alpha X + \delta H + N_Y$, where H is hidden.

  Introduce Z independent of H, dependent on X, and affects Y only via X.

  Then $X = \beta Z + \gamma H + N_X = \beta Z +$ some noise .

  Estimate β by regressing X on Z.

  Go back and substitute:

  $$Y = \alpha X + \delta H + N_Y = \alpha(\beta Z + \gamma H + N_X) + \delta H + N_y = \alpha(\beta Z) + (\alpha\gamma + \delta)H + N_Y.$$

  Now, regressing Y on βZ gives the estimate for α.

10 Pages 197-211

- Multivariate time series: causal structures need to be consistent with the time order. So, causal influence (arrows) can never go from the future to the past. This makes it easier to recover the causal DAG (faithfulness is not needed anymore, minimality+Markov are sufficient, and there is no need to restrict the function class either).
  - Two types of graphs for time series: full time graph (nodes are each of the features at each time step... potentially infinite number of nodes) and summary graph (one node per feature). The summary graph may contain cycles, but the full time graph is a DAG.
  - Two DAGs are Markov equivalent iff their skeleton and v-structures are the same. For full time graphs, if they are Markov equivalent, they are the same (in absence of instantaneous effect) because, in addition to the general case, we also know the direction of arrows. If there are instantaneous effects, the direction of those is the only thing that might be different between two Markov equivalent graphs. But still, when instantaneous effect have different directions, v-structures change. So, from Markov condition and faithfulness, and from conditional independences, we can **uniquely** identify the full time graph (whereas before, those only gave us the equivalence class), assuming it is DAG.
  - Faithfulness is not strictly needed, minimality is enough. This is because time gives the possible direction of arrows, so each arrow in the summary graph comes from conditional dependence given the past, and if two nodes are conditionally independent given the past, then there is no arrow in the summary graph.
  - Dynamic Causal Modeling: modeling with differential equation. $\frac{d}{dt}z = F(z, u, \theta)$ models the variation of z. The model can be approximated to linear, having matrix coefficients for mutual influences of z_i, mutual influence of u_i, and influence of u, where u is a vector of perturbations. DCM has been criticized because the number of model parameters explodes with the sizes of z and u, so identification becomes impossible from empirical data (too much data needed). Also, from experiments with simulated brain connections, a large fraction of wrong models obtained evidence by DCM. (so maybe DCM is irrelevant, presented in the chapter only for the sake of completeness? Or I am not understanding it at all.)
- Granger causality: in a time series, the cause happens before the effect, but also, the cause contains information that is useful to predict the value of the effect. So, one conditional independence test is enough to decide whether there is an arrow or not. However, if causal sufficiency is violated, then the results can not be interpreted causally.

References

[1] Peters, J., Janzing, D., Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* The MIT Press. 2017.