Causal Falling Rule Lists

Fulton Wang[1], **Cynthia Rudin**[1]
[1]MIT

Abstract

A causal falling rule list (CFRL) is a sequence of if-then rules that specifies heterogeneous treatment effects, where (i) the order of rules determines the treatment effect subgroup a subject belongs to, and (ii) the treatment effect decreases monotonically down the list. A given CFRL parameterizes a hierarchical bayesian regression model in which the treatment effects are incorporated as parameters, and assumed constant within model-specific subgroups. We formulate the search for the CFRL best supported by the data as a Bayesian model selection problem, where we perform a search over the space of CFRL models, and approximate the evidence for a given CFRL model using standard variational techniques. We apply CFRL to a census wage dataset to identify subgroups of differing wage inequalities between men and women.

Introduction

In identifying heterogeneous treatment effects, the end goal is often to rank subgroups by the treatment effects within, so that those for which a treatment is most effective can be treated first. This segmentation of data into regions of differential treatment effect has been of recent interest in social science, medical, and marketing domains (Brodersen et al. 2015; Sun et al. 2014; Cai et al. 2011; Foster, Taylor, and Ruberg 2011; Taddy et al. 2014), precisely so that the relevant subgroups can be given priority treatment. For example, a drug can be given to the patient group for whom it is most effective, or an ad can be shown to the audience most likely to be swayed by it. Methods have used tree structures (Su, Kang, and Levine 2012; Su and Nickerson 2009; Athey and Imbens 2015; Beygelzimer and Langford 2009) to form such treatment effect subgroups. Rule trees have the benefit of being interpretable and transparent in defining the partitions, and (potentially) sparse in the number of partitions. However, tree based methods suffer from two drawbacks: (i) their training is based on greedy splitting criteria, and (ii) given a partitioning tree, it is still cognitively demanding to perform the downstream decision-making task of ranking subgroups by treatment effect and identifying the logical combination of rules defining each.

To address these shortcomings of past tree-based methods for estimating heterogeneous treatment effects, we introduce *causal falling rule lists*. A causal falling rule list (CFRL) is a Bayesian model parameterized by a sequence of if-then rules such that (i) the sequence of rules determines which treatment effect subgroup a subject belongs to, and (ii) the treatment effect for each subgroup *decreases monotonically* as one moves down the list. For example, a CFRL might say that if a person is below 60 years, then they are in the highest treatment effect subgroup, such that administering a drug will result in a 20 unit increase in good cholesterol levels. Otherwise, if they are regular exercisers, then taking the drug will result in a 15 unit increase in cholesterol level. Finally, if they satisfy neither of these rules, they are in the default treatment subgroup, such that the drug will result in only a 2 unit increase. Thus, the hallmark of a given CFRL is that the treatment effect is modelled as being *constant* within the single subgroup partition that the given CFRL defines and "falling" along the subgroups.

The special structure of a CFRL addresses the shortcomings of tree-based methods. Firstly, non-greedy training procedures become feasible when the search space is reduced from that of trees to lists. Secondly, the monotonicity constraint over treatment effects directly ranks the treatment effect subgroups in the order they should be targeted.

Past work informs our approach to learning a CFRL model. (Su and Nickerson 2009; Su, Wang, and Fan 2004; Zeileis, Hothorn, and Hornik 2008) use a model-based node splitting criteria in building a tree that identifies heterogeneity of some quantity of interest based on covariates. In Su and Nickerson, for example, this quantity is treatment effect in a randomized trial, and the quality of a split is determined by a t-test that the difference in control and treated outcome between splits is 0 under a model where treated and control outcomes are assumed constant within each split. Such methods have the benefit of being explicit about the assumptions underlying the model training procedure, but one drawback is that they are not directly optimizing some global criteria like likelihood. Furthermore, the training depends on complexity penalties which must be chosen via cross validation. Imai and Ratkovic formulate the estimation of heterogeneous treatment effects as a model selection problem. In particular, they use variable selection on treatment/covariate interaction features in a SVM classification model to iden-

tify a parsimonious set of interactions between treatment and covariates that explain the (binary outcome) data. One drawback of this approach is that the number of features identifying each treatment effect subgroup is the number of selected treatment/covariate interaction features, which will generally be too large to be interpretable, and the number of subgroups will be large - exponential in that number.

Our approach will be model based, but we instead assume a single model for explaining *all* of the data, instead of just for node-splitting purposes. In particular, a given CFRL, as previously mentioned, is parameterized by a single partition of the data into ordered subgroups such that within each subgroup, the likelihood assumes the treatment effect to be constant and "falling". Instead of searching for the CFRL for which a score incorporating the corresponding maximum likelihood and some complexity penalty is highest, we choose to formulate the search for the "best" CFRL as a *Bayesian* model selection problem. Thus for a given CFRL, we place a prior over its parameters and choose the CFRL for which the evidence, the likelihood averaged over the parameter prior, is highest. This approach uses Bayesian Occam's Razor to penalize overly complex models, and avoids having to perform cross validation to tune the complexity penalty parameters. The model search process will identify heterogenous treatment effects by identifying the model under which the assumption of "falling" constant treatment effects across the subgroups of the given model are most likely. We believe a Bayesian framework is appealing because it provides in a single unified framework a way to choose between different models and provide uncertainty estimates conditional on the choice of a single model.

Thus, a CFRL model is a Bayesian hierarchical regression model, where the sequence of rules in it determines the grouping of data, and a regression model for each subgroup relates controlling covariates and treatment status to outcome. Our model is hierarchical because the regression coefficient vectors for each subgroup are allowed to vary, but are generated by a common unknown distribution and thus dependent. The amount of sharing of strength between the grouped parameters can be specified via a prior on variance parameters. The model search process consists of simulated annealing over model space, where local changes are proposed to sequence of rules parameterizing the current model.

To make the model selection process tractable, we make two approximations. First, we restrict the model space to CFRL models consisting of rules drawing from a premined set such as that from a rule miner like FPGrowth (Borgelt 2005). The rules of the rule set are chosen so that they have non-negligible support and are interpretable, and its total size can be limited as necessary for computational tractability. Second, we use a standard variational approximation for the evidence of a given CFRL model, as this calculation is required at each simulated annealing step. Thus, the model selection process is similar to that of Bayesian Hierarchical Mixture of Experts (Bishop and Svenskn 2002), but with hard partitioning of data and dependence between experts.

We believe the biggest contribution of our model is its interpretability - the clear manner in which it communicates its treatment effect estimates. Interpretability is highly context dependent (Kodratoff 1994; Pazzani 2000; Martens et al. 2011); no matter how one measures it in one domain, it can be different in the next domain. However in a setting where one is deciding which subgroup to treat, CFRL has the potential to directly guide the decision-making process. In addition to stratifying and ranking subjects by treatment effect, a CFRL is in practice, as sparse as one desires. Since it automatically stratifies subjects by risk in the order used for decision making, one can choose to look at as much of the list as they need to make a treatment decision. If one only wishes to treat those for which the treatment is most effective, they need only look at the top few rules, and check whether the subject obeys any of the top clauses.

Our paper is organized as follows: We first describe the generative process for our model, and then detail the simulated annealing model search process and the variational evidence approximation. We then show that given simulated data, this procedure can recover the true model as well as the parameters for the model. Finally, we apply CFRL to a census wage dataset to characterize the heterogeneous effect of being male on wages.

Model

Notation

In this work, we assume a binary treatment and a dataset of N units indexed by (n), who each have K-dimensional covariate vectors $x^{(n)} \in \mathbb{R}^K$. We use the Rubin potential outcomes framework (Rubin 1974), with potential outcomes $Y_1^{(n)}, Y_0^{(n)} \in \mathbb{R}$ under treatment and control, respectively, and treatment assignment indicator $T^{(n)} \in \{0, 1\}$. x, Y_1, Y_0, T will refer to the set of covariates, potential outcomes, and treatment assignments for all N units in the training data, collectively. We denote the observable outcome for the n-th unit as

$$Y^{(n)} = T^{(n)} Y_1^{(n)} + (1 - T^{(n)}) Y_0^{(n)}, \tag{1}$$

and the unobservable outcome as

$$\bar{Y}^{(n)} = (1 - T^{(n)}) Y_1^{(n)} + T^{(n)} Y_0^{(n)}. \tag{2}$$

Similarly Y and $\bar{Y}$ to refer to the observable and unobservable outcomes for the entire data. According to the Bayesian approach we will assume a distribution $P(Y_1, Y_0, T, \theta | x)$, where θ refers to unknown parameters of our model. Given observed data, the goal will be to, for a test sample indexed by $(*)$, under our model, present the posterior of $Y_1^{(*)} - Y_0^{(*)} | x^{(*)}$. This distribution will depend only on the posterior of θ, whose inference will be our main focus.

Assumptions

We assume that all covariates x and observable outcomes Y are available. Secondly we assume the Bayesian version of conditional ignorability (Rubin 1978), that under the assumed model $P(Y_1, Y_0, T, \theta | x)$,

$$T \perp\!\!\!\perp \theta, Y_1, Y_0 | x \tag{3}$$

Under this assumption, $P(Y_1, Y_0, T, \theta | x) = P(Y_1, Y_0, \theta | x)$ regardless of what $P(T|x)$ is. Thus, the modeller need only provide a model of $P(Y_1, Y_0, \theta | x)$, and we provide this model under the factorization $P(Y_1, Y_0 | \theta, x) P(\theta | x)$.

Parameterization

Accordingly, a CFRL is a Bayesian model of $Y_1^{(n)}, Y_0^{(n)}$ given $x^{(n)}$.A given CFRL is parameterized by the length and sequence of rules in it (and subsequent hyperparameters, specified later):

$$L \in \mathcal{Z}^+ \qquad \text{(length of list)} \tag{4}$$

$$c^{(l)}(\cdot) \in C \quad \text{for } l = 1, \ldots, L \qquad \text{(rules in list)} \tag{5}$$

where C represents the space of rules, namely that of boolean functions on feature space $\mathbb{R}^K$. These L rules partition the feature space into L regions within which the treatment effects are assumed constant. We notate a model by its rule list $\{c^{(l)}(\cdot)\}$ or $M; \{c^{(l)}(\cdot)\}$, omitting dependence on hyperparameters when appropriate.

Generative Process

Under a model with rule sequence $\{c^{(l)}(\cdot)\}$, which we will denote $M; \{c^{(l)}(\cdot)\}$, the subjects are assigned treatment effect subgroup $z^{(n)} \in \{1, \ldots, L\}$ according to the logic of a decision list:

$$z^{(n)} = \min(l; c^{(l)}(x^{(n)}) = 1, l = 1, \ldots, L). \tag{6}$$

We will always assume that the last rule, $c^{(L)}(\cdot)$, is a *default rule* that always returns true, so that this $\min$ is well defined.

A separate regression model within each subgroup controls for confounding covariates and models the impact of receiving the treatment on outcome, giving the likelihood:

$$Y_1^{(n)} | x^{(n)} \sim \mathcal{N}(D^{(z^{(n)})} + B^{(z^{(n)})\prime} x^{(n)}, \tfrac{1}{\lambda^{(z^{(n)})}}), \tag{7}$$

$$Y_0^{(n)} | x^{(n)} \sim \mathcal{N}(B^{(z^{(n)})\prime} x_n, \tfrac{1}{\lambda^{(z^{(n)})}}) \tag{8}$$

depending on parameters

$$D^{(l)} \in \mathbb{R} \qquad \text{(subgroup treatment effect)} \tag{9}$$

$$B^{(l)} \in \mathbb{R}^K \qquad \text{(subgroup regression coefficient)} \tag{10}$$

$$\lambda^{(l)} \in \mathbb{R}^+ \qquad \text{(subgroup noise precision)} \tag{11}$$

for $l = 1, \ldots, L$, and under the constraint that the treatment effects decrease down the list:

$$D^{(l)} > D^{(l-1)} \qquad \text{for } l = 1, \ldots, L-1. \tag{12}$$

The pattern of missing data can be ignored due to conditional ignorability, so only observed data need be modelled.

Prior

The joint prior over subgroup treatment effects $D^{(1)}, \ldots, D^{(L)}$ must respect the monotonicity constraints of Equation (12). Thus, we perform the reparameterization

$$D^{(l)} = \sum_{l'=L}^{l} \delta^{(l')} \tag{13}$$

and place uniform priors with only support over the positive reals on (all but one of) the newly introduced $\delta^{(l)}$:

$$\delta^{(l)} \sim \text{Uniform}(0, s_0) \text{ for } l = 1, \ldots, L-1, \tag{14}$$

$$\delta^{(L)} \sim \text{Uniform}(r_0, s_0), \tag{15}$$

with $s_0 \geq 0$. Thus we enforce the monotonicity of Equation (12) as a "hard" constraint that will still be true in the posterior. For example, we know that in the posterior, $E[D^{(l)}] > E[D^{(l+1)}]$.

We assume each $B^{(l)}$ is written as the concatenation $B^{(l)} = [B_h^{(l)} B_i^{(l)}]$ with $B_h^{(l)} \in \mathbb{R}^{K_h}$, $B_i^{(l)} \in \mathbb{R}^{K_i}$, $K_h + K_i = K$, where strength is shared between the $B_h^{(l)}$ through an hierarchical prior, and the $B_i^{(l)}$ are assumed independent in the prior:

$$\mathbb{R}^+ \ni \tau \sim \text{Wishart}(v_0, w_0) \tag{16}$$

$$\mathbb{R}^{K_h} \ni m \sim \mathcal{N}(\mathbf{0}_{K_h}, (c_0 I_{K_h})^{-1}) \tag{17}$$

$$B_h^{(l)} \sim \mathcal{N}(m, (\tau I_{K_h})^{-1}) \tag{18}$$

and

$$B_i^{(l)} \sim \mathcal{N}(\mathbf{0}_{K_i}, (u_0 I_{K_i})^{-1}), \tag{19}$$

where $\mathbf{0}_K$ denotes the K-dimensional 0-vector, I_K denotes the K-dimensional identity matrix, and the Wishart distribution is 1-dimensional (a reparameterized Gamma).

Finally, we let

$$\lambda^{(l)} \sim \text{Gamma}(\alpha_0, \beta_0) \tag{20}$$

so that the complete set of hyperparameters are

$$s_0, v_0, w_0, c_0, u_0, \alpha_0, \beta_0 \in \mathbb{R}^+, \tag{21}$$

$$r_0 \in \mathbb{R}. \tag{22}$$

Remarks on Prior

Firstly, we comment that is necessary for some features to have corresponding independent regression coefficients across treatment subgroups. For example, the bias for a subgroup (the regression coefficient for a feature that is uniformly 1) should be independent across subgroups; if the bias is drawn in one direction by a hierarchical prior, then the treatment effect will be drawn in the other direction to compensate.

Secondly, the prior on τ, the precision parameter controlling the dependency between subgroup regression coefficients can be set to one's preferences. For example, one might look favorably upon a model in which variation in the regression parameters across subgroups allows the data to be explained very well (subject to the "falling" constraint on treatment effects), and place most prior mass of τ close to 0. However, one might believe that interactions between treatment effect subgroup membership and confounder effects is not plausible, and encourage strong sharing between the regression parameters. Issues raised in (Gelman and others 2006) do not arise because the number of observations of the distribution parameterized by τ is high (linear in the number of features).

Model Selection

As mentioned in the introduction, we take a Bayesian approach to choosing the CFRL that best explains the data - we want to find the model for which the evidence is highest. Thus, the model selection process consists of two tasks: (i) calculating the evidence of a given model, and (ii) maximizing the evidence over the space of models, which we now describe in turn.

Evidence Approximation

For a model $M; \{c^{(l)}(\cdot)\}$, we must calculate the evidence

$$p(Y; M) = \int_\theta p(Y|\theta; M)p(\theta; M)d\theta, \text{ where} \quad (23)$$

$$\theta = \{\vec{B}, \vec{\delta}, \vec{\lambda}, m, \tau\} \quad (24)$$

are the latent parameters, where $\vec{B}$ denotes the L parameters $\{B^{(l)}\}$ (likewise for $\vec{\delta}, \vec{\lambda}$), and

$$\mathcal{H} = \{s_0, v_0, w_0, c_0, u_0, \alpha_0, \beta_0, r_0\}. \quad (25)$$

are the hyperparameters of the model.

As this integral is not analytically available, and computationally infeasible to calculate using sampling, we approximate it using a standard variational approach.

Variational Inference In variational inference, a key observation is that for *any* distribution $q(\cdot)$ over θ,

$$\log p(Y) = L(q(\theta)) + \mathcal{KL}(q(\theta)||p(\theta|Y)) \quad (26)$$

$$> L(q(\theta)), \text{ where} \quad (27)$$

$$L(q(\theta)) = \int_\theta q(\theta) \log \frac{p(\theta|Y)}{q(\theta)} d\theta. \quad (28)$$

and equality holding only if $KL(q(\theta)||p(\theta|Y)) = 0$. Note that all distributions depend on M, but we suppress that dependency when convenient.

A lower bound on $\log p(Y)$ can thus be obtained by choosing a family of distributions $\mathcal{Q}$ on θ, and calculating

$$L^*(M; \mathcal{Q}) = \max_{q \in \mathcal{Q}} L(q(\theta); M). \quad (29)$$

To make this maximization tractable, in mean-field inference, a partition of θ, $\{\theta_i\}$ is chosen, and $\mathcal{Q}$ is chosen to be the family of distributions that factorizes over $\{\theta_i\}$:

$$\mathcal{Q} := \{q(\theta) = \{\Pi_i q_i(\theta_i); q_i \in \mathcal{P}_{\theta_i}\}. \quad (30)$$

For an optimal $q^* \in \text{argmax}_{q \in \mathcal{Q}} L(q(\theta); M)$, it can be shown that the following conditions are satisfied: For each i,

$$q_i^*(\theta_i) \propto E_{q_{\theta_{-i}}}[\log p(\theta_i, \theta_{\setminus i}|Y)] \quad (31)$$

where $E_{q_{\theta_{-i}}}[\cdot]$ is an expectation over the distribution $q_{-i}(\theta_{-i}) = \Pi_{j \neq i} q_j(\theta_j)$. These conditions imply coordinate ascent *variational updates* of maximizing $q_i(\theta_i)$ holding $\{q_j(\theta_j)\}_{j \neq i}$ fixed. Furthermore, for distributions $p(\theta|Y)$ where the complete conditional distributions $p(\theta_i|\theta_{-i}, Y)$ are in the exponential family, $q_i^*(\theta_i)$ is in the same family.

Variational Updates The family $\mathcal{Q}$ over which we calculate $L^*(M; \mathcal{Q})$ is one that factorizes as:

$$q(\vec{B}, \vec{\delta}, \vec{\lambda}, m, \tau) = q_{\vec{B}}(\vec{B}) q_{\vec{\delta}}(\vec{\delta}) q_{\vec{\lambda}}(\vec{\lambda}) q_{m,\tau}(m, \tau) \quad (32)$$

Because of the conditional conjugacies in our model, $\text{argmax}_{q \in \mathcal{Q}} L(q(\theta))$ lies in the distribution family $\mathcal{Q}' \subset \mathcal{Q}$ which is parameterized by *variational parameters*

$$\xi = \Big\{\{\mu_{B^{(l)}}, \Sigma_{B^{(l)}}\}_l, \{\alpha_{\lambda^{(l)}}, \beta_{\lambda^{(l)}}\}_l, \mu_{\vec{\delta}}, \Sigma_{\vec{\delta}}, v_\tau, w_\tau, \mu_m, c_m\Big\} :$$

$$q_{\vec{B},\vec{\delta},\vec{\lambda},m,\tau}(\vec{B}, \vec{\delta}, \vec{\lambda}, m, \tau; \xi) = \quad (33)$$

$$q_{\vec{\delta}}(\vec{\delta}; \mu_{\vec{\delta}}, \Sigma_{\vec{\delta}})\ q_{\tau,m}(\cdot; v_\tau, w_\tau, \mu_m, c_m)$$

$$\Pi_l\ q_{B^{(l)}}(B^{(l)}; \mu_{B^{(l)}}, \Sigma_{B^{(l)}})\ q_{\lambda^{(l)}}(\lambda^{(l)}; \alpha_{\lambda^{(l)}}, \beta_{\lambda^{(l)}})$$

with

$$q_{\vec{\delta}}(\cdot; \mu_{\vec{\delta}}, \Sigma_{\vec{\delta}}) = p^{\mathcal{N}(r_0,s_0)}(\cdot; \mu_{\vec{\delta}}, \Sigma_{\vec{\delta}})$$

$$q_{B^{(l)}}(\cdot; \mu_{B^{(l)}}, \Sigma_{B^{(l)}}) = p^{\mathcal{N}}(\cdot; \mu_{B^{(l)}}, \Sigma_{B^{(l)}})$$

$$q_{\lambda^{(l)}}(\cdot; \alpha_{\lambda^{(l)}}, \beta_{\lambda^{(l)}}) = p^{\gamma}(\cdot; \alpha_{\lambda^{(l)}}, \beta_{\lambda^{(l)}})$$

$$q_{\tau,m}(\cdot; v_\tau, w_\tau, \mu_m, c_m) = p^W(\tau; v_\tau, w_\tau)\ p^{\mathcal{N}}(m; \mu_m, (c_m \tau I_{D_h})^{-1}).$$

Here, $p^W(\cdot)$ denotes a 1-dimensional Wishart distribution and the distribution of $\vec{\delta}$, $p^{\mathcal{N}(r_0,s_0)}(\cdot; \mu, \Sigma^{-1})$ is a normal distribution *truncated* so that $\delta^{(1)}, \ldots, \delta^{(L-1)}$ are truncated to the range $(0, s_0)$ and $\delta^{(L)}$ is truncated to the range (r_0, s_0). The truncation is due to the prior support of $\vec{\delta}$.

The coordinate ascent updates of Equation (31) are then over ξ, and can be derived to be as follows:

update for $q_{\vec{\delta}}^*$:

$$\Sigma_{\vec{\delta}}^{-1} \leftarrow \sum_l (E_{q_{\lambda^{(l)}}}[\lambda^{(l)}] \sum_n z^{(nl)} T^{(n)}) k_l k_l'$$

$$\mu_{\vec{\delta}} \leftarrow \Sigma_{\vec{\delta}} \sum_l \Big(E_{q_{\lambda^{(l)}}} \sum_n z^{(nl)} T^{(n)} (Y^{(n)} - E_{q_{B^{(l)}}}[B^{(l)}]' x^{(n)}) \Big) k_l$$

update for $q_{B^{(l)}}^*$:

$$\Sigma_{B^{(l)}}^{-1} \leftarrow \text{diag}[E_{q_{m,\tau}}[\tau] I_{D_h}, u_0 I_{K_i}] + E_{q_{\lambda^{(l)}}}[\lambda^{(l)}] \sum_n z^{(nl)} x^{(n)} x^{(n)\prime}$$

$$\mu_{B^{(l)}} \leftarrow \Sigma_{B^{(l)}} \Big(\text{diag}[E_{q_{m,\tau}}[\tau m] I_{K_h}, 0_{K_i}] + E_{q_{\lambda^{(l)}}}[\lambda^{(l)}] \sum_n z^{(nl)} (Y^{(n)} - T^{(n)} k_l' E_{q_{\vec{\delta}}}[\vec{\delta}] x^{(n)}) \Big)$$

update for $q_{\tau,m}^*$:

$$v_\tau \leftarrow v_0 + K_h L$$

$$c_m \leftarrow c_0 + L$$

$$\mu_m \leftarrow \tfrac{1}{c_0 + L} \sum_l E_{q_{B^{(l)}}}[B_h^{(l)} B_h^{(l)\prime}] + \tfrac{c_0}{c_0 + L} m_0$$

$$w_\tau \leftarrow (w_0^{-1} + \sum_l E_{q_{B^{(l)}}}[B_h^{(l)\prime} B_h^{(l)}])^{-1}$$

update for $q_{\lambda^{(l)}}^*$:

$$\alpha_{\lambda^{(l)}} \leftarrow \alpha_0 + \tfrac{1}{2} \sum_n z^{(nl)}$$

$$\beta_{\lambda^{(l)}} \leftarrow \beta_0 + \tfrac{1}{2} \sum_n z^{(nl)} E_{q_{\vec{\delta}} q_{B^{(l)}}}[(Y^{(n)} - (T^{(n)} k_l' \vec{\delta} + B^{(l)\prime} x^{(n)}))^2]$$

where we have defined $z^{(nl)} = \mathbb{1}[z^{(n)} = l]$ and k_l is the length L vector whose first $l - 1$ entries are 0, 1 otherwise, and $\text{diag}[\ldots]$ is a block diagonal matrix with the specified blocks.

All involved expectations depend on the current value of variational parameters (suppressed in notation for conciseness), and are available in closed form with the exception of expections over $q_{\vec{\delta}}(\cdot)$, a truncated normal distribution. For those we used the R package from (Wilhelm 2012). We note that we chose $\mathcal{Q}$ to consist of distributions that factorize *separately* over $\vec{\delta}$ and $\{B^{(l)}\}$ though not

doing so would still preserve conjugacy. However in this case, $q_{\vec{\delta},\{B^{(l)}\}}$ would be a high dimensional truncated normal distribution, whose first and second moments, required for the variational updates, were not tractable to calculate. Finally, applying these updates until convergence approximates $q^* = \text{argmax}_{q \in \mathcal{Q}} L(q(\theta); M)$. We then calculate $L(q^*(\theta))$ to obtain a lower bound on $p(Y; M)$. (We omit details of this straightforward but tedious calculation.)

Model Search

We now detail the model space over which we maximize the model evidence, and the search procedure for doing so.

Model Space As mentioned earlier, for computational feasibility and interpretability purposes, we only perform model selection over models $M; \{c^{(l)}(\cdot)\}$ where each rule of the rule list is assumed to come from a pre-mined set of boolean functions C returned by a frequent item-set mining algorithm. For this particular work, we used FPGrowth (Borgelt 2005), whose input is a binary dataset where each x is a boolean vector, and whose output is a set of subsets of the features of the dataset. For example, x_2 might indicate the presence of diabetes, and x_{15} might indicate the presence of hypertension, and a boolean function returned by FPGrowth might return 1 for patients who have diabetes and hypertension. The number of clauses in a rule can be limited for interpretability. Finally, we fix the hyperparameters to be the same for each considered model.

Simulated Annealing We use simulated annealing over model space to maximize model evidence. Given a objective function $E(s)$ over discrete search space S, a function specifying the set of neighbors of a state $N(s)$, and a temperature schedule function over time steps, $T(t)$, a simulated annealing procedure is a discrete time, discrete state Markov Chain $\{s_t\}$ where at time t, given the current state s_t, the next state s_{t+1} is chosen by first randomly selecting a proposal $\grave{s}$ from the set $N(s)$, and setting $s_{t+1} = \grave{s}$ with probability $\min(1, \exp(-\frac{E(\grave{s})-E(s)}{T(t)}))$, and $s_{t+1} = s_t$ otherwise.

S in this setting is the set of sequences of rules $\{c^{(l)}(\cdot)\}$ with $c^{(l)}(\cdot) \in C$, the premined rule set, and accordingly we let $E[\{c^{(l)}(\cdot)\}] = L^*(\{c^{(l)}(\cdot)\}; \mathcal{Q})$ from Equation (29). Model dependence on hyperparameters is suppressed as they are assumed fixed, and so a model M is parameterized solely by its rule list. We simultaneously define the set of neighbors and the process by which to randomly choose a neighbor through the following random procedure that alters the current rule list $\{c^{(l)}(\cdot)\}_{l=1}^{L}$ to produce a new rule list $\{\grave{c}^{(l)}(\cdot)\}_{l=1}^{\grave{L}}$ (The new list's length may be different):

Choose uniformly at random one of the following 4 operations to apply to the current rule list, $\{c^{(l)}(\cdot)\}_{l=1}^{L}$:

1. SWAP: Select $i \neq j$ uniformly from $1, \ldots, L-1$, and swap the rules at those 2 positions, letting $\grave{c}^{(i)}(\cdot) \leftarrow c^{(j)}(\cdot)$ and $\grave{c}^{(j)}(\cdot) \leftarrow c^{(i)}(\cdot)$.
2. REPLACE: Select i uniformly from $1, \ldots, L-1$, choose a rule $c(\cdot)$ from C not already in the rule list, and set $\grave{c}^{(i)}(\cdot) \leftarrow c(\cdot)$.
3. ADD: Choose one of the L possible insertion points uniformly at random, draw a rule $c(\cdot)$ from C, and insert it at the chosen insertion point, so that $\grave{L} \leftarrow L + 1$.
4. REMOVE: Choose i uniformly at random from $1, \ldots, L-1$, and remove $c^{(i)}(\cdot)$ from the current rule list, so that $\grave{L} \leftarrow L - 1$.

Note that since the last rule is always assumed to be the default rule, we never modify the last rule in the list.

Simulation Studies

We show that for simulated data generated by a known CFRL model, our simulated annealing procedure that maximizes the evidence over model space, with high probability, recovers the model.

Given observations with arbitrary features, and a collection of rules on those features, we can construct a binary matrix where the rows represent observations and the columns represent rules, and the entry is 1 if the rule applies to that observation and 0 otherwise. We need only simulate this binary matrix to represent the observations without losing generality. For our simulations, we generated independent binary rule sets with 100 rules by setting each feature value to 1 independently with probability 0.25.

Then, for each N, we performed the following procedure 20 times: We generated the random rule matrix, generated a random CFRL of size 6 by selecting 5 rules at random plus the default rule, and, assuming 10 confounding features, for $l = 1, \ldots, 6$, generated pararameters $B^{(l)} \sim \mathcal{N}(0_{10}, I_{10})$ and set $\lambda^{(l)} = 1$. For each $n = 1, \ldots, N$ we generate $x^{(n)} \sim \mathcal{N}(0_{10}, I_{10})$, $T^{(n)}$ uniformly from $\{0, 1\}$ and simulate $Y^{(n)}$ according to Equation (7), (8) and (1) to obtain an independent dataset of size N. We then run the simulated annealing procedure to obtain an estimate of the true model. We calculated the edit distance of the rule list obtained from simulated annealing to the true rule list model, and display in Figure 2 the average distance over these 20 replicates, as N varies. Note that the maximum possible edit distance is 5, as the 6th rule is always the default rule. Regarding priors, we placed an hierarchical prior over all regression coefficients, used the same hyperparameters as in the analysis of wage data, detailed in the next section.

Application to Wage Data

We apply CFRL to a dataset that contains the hourly wage of individuals to assess the treatment effect of gender on hourly wage. We view being male as the treatment. This dataset was collected in 1995 through the US Census' Current Population Survey (US Census 1995) and after removing individuals who were unemployed or for whom salary data was not available, the dataset retains the salary and gender of 7548 individuals, along with 15 covariates. The covariates included mostly categorical features such as industry, marital status, union status, education level, as well as 2 scalars: age and weeks worked. Because the FPGrowth rule miner we use accepts binary features only, we discretize the scalar features and group the levels of categorical features manually to obtain $K = 54$ binary features after adding 1 addi-

	Conditions		Support	Effect	Match	non-truncated
IF	Occup=prof. specialty AND race=white	THEN treatment effect is:	579	$5.49	$4.01	$4.39
ELSE IF	Occup=factory AND union=no	THEN treatment effect is:	531	$3.93	$2.14	$2.13
ELSE IF	Occup=sales AND householder=false	THEN treatment effect is:	492	$2.32	$1.07	$1.38
ELSE IF	Industry=trade AND householder=false	THEN treatment effect is:	649	$2.08	$0.40	$1.21
ELSE IF	govt employer=false AND no college educ.	THEN treatment effect is:	3939	$1.86	$2.73	$1.12
ELSE IF	Industry=education	THEN treatment effect is:	255	$1.11	$1.36	$0.76
ELSE		treatment effect is:	1103	$0.55	$0.75	$0.42

Figure 1: Causal falling rule list for treatment effect of being male on hourly wage.

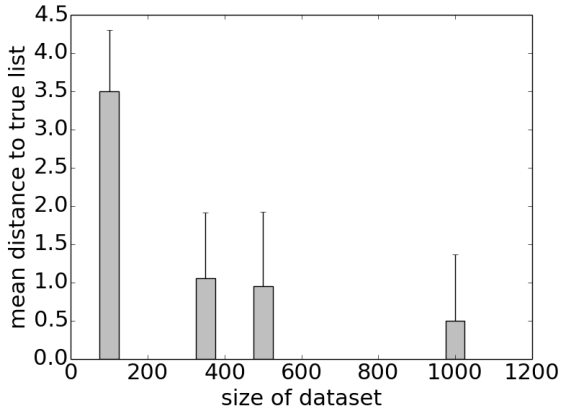

Figure 2: Mean distance to true list decreases with increasing sample size.

tional bias feature, so that each $x^{(n)} \in \{0,1\}^{54}$. We then mine all rules with a support of at least 5% and at most 2 clauses, to obtain a set of 561 rules, C. The range of hourly wages had a mode of around \$8, with few hourly wages above \$25.

Regarding priors, we placed an hierarchical prior on all except the bias feature so that $K_h = 53$ and $K_i = 1$. We let the hierarchical prior hyperparameters $v_0 = 2, w_0 = 0.5, c_0 = 10$, and let $u_0 = \frac{1}{25^2}$. With this prior (verified through simulations), for both features d with hierarchical and independent priors, prior values of $B_d^{(l)}$ are generally in $[-25, 25]$, which is reasonable given the range of salaries. For features with hierarchical priors, for l, l', $B_d^{(l)}$ and $B_d^{(l)\prime}$ are moderatedly correlated. We let $r_0 = -25$, $s_0 = 25$.

With this prior, we ran simulated annealing for 5000 steps, with a constant temperature of 1, and initializing with a random rule list of length 3. We display in Figure 1 the rules of the model with the highest evidence. The posterior predictive distribution of treatment effect for a test sample, $Y_1^{(*)} - Y_0^{(*)} | x^{(*)}$, for which neither potential outcome is observed, is simply $D^{z^{(*)}}$, where $z^{(*)}$ denotes the treatment effect subgroup the test unit belongs to based on their features and the rules in the rule list according to Equation (6). Thus, we display in the column "Effect" the variational posterior mean of each $D^{(l)}$, and in Figure 3, we plot the corresponding distributions. The posterior means are "falling" due to the hard prior constraint of Equation (12). To compare the variational posterior to the true posterior, we implemented Gibbs sampling for the model, and show the posteriors of each $D^{(l)}$ from 7500 Gibbs steps (with 2500 burn-in steps) in the bottom of Figure 3.

Our model selection procedure has given us a ordered partition of the data for which the evidence is high, under a "falling" prior over treatment effects, and posterior inference for the model gives the corresponding treatment effect estimates for each partition, which are guaranteed to be "falling". However given the partition, we can of course for comparison estimate via other methods the subgroup treatment effects. Thus, we show the treatment effect estimates obtained via propensity score matching within each subgroup using the Matching R package (Sekhon 2011) in the column "Match". Secondly, we fit a non-"falling" version of our model, where the $\delta^{(l)}$ parameters are given a uniform prior with equal positive and negative support, and show the posterior mean $D^{(l)}$ parameters in the "non-truncated" column. Neither of these estimates are guaranteed to be "falling", though they may be more reasonable depending on the user's purposes.

One notices that the mean posterior treatment effects of our model are inflated relative to those from matching. However, this is inevitable due to the monotonicity constraint on the joint prior over subgroups treatment effects. Any sample from the prior will have $D^{(l)} > D^{(l+1)}$. Thus if the prior of $D^{(L)}$ corresponding to the bottom of the list is "unbiased" in that it has a mean of 0, it is inevitable that the prior means of $D^{(l)}$ will become higher and more "biased" as one moves towards the top of the list (as l decreases). The effect of the prior will decrease as the amount of data increases, but we mention a potential remedy for this bias in the conclusion. Also note that the estimates from the non-"falling" version of our model, whose priors do not have the monotonicity-induced bias, are more similar to those from matching than those from the original CFRL model.

Regarding the rules in the model, "occup" is short for occupation, and examples of professional specialty occupations are engineers, lawyers, architects; "union=no" means the person is not in a union; "householder=false" means one is not the primary homeowner (and could either be the spouse of one or a child.)

Presented with a CFRL, due to its easy to understand structure, one (we hope) is inclined to see whether it matches common sense. The second rule indicates a high male salary

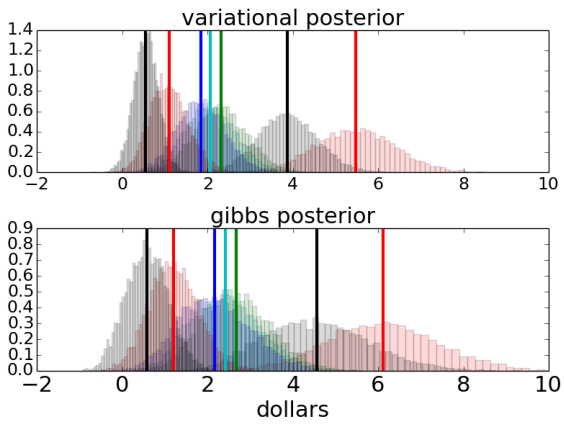

Figure 3: Posterior treatment effects for each subgroup as obtained by variational inference and Gibbs sampling. Vertical bars indicate location the mean of the respective distributions. Due to the "falling" constraint, the rightmost distribution corresponds to the treatment effect for top subgroup in the list, the 2nd-rightmost corresponds to second from the top subgroup, and so forth.

relative to equivalent women among factory workers, but only among those not in unions. This makes intuitive sense, and we also note this interaction is enabled by the fact that rules can contain conjunction of features. The fifth rule captures captures the largest number of people, and suggests that the wage discrepancy may be less amongst government employees, which again makes some intuitive sense.

Conclusion

We have created a new model that estimates treatment effects, and has a form that immediately informs the decision making process of whom to apply a treatment to. We envision a CFRL to be used as a decision guide in resource-constrained environments. Furthermore, the easy to understand and interpretable structure of a CFRL allows one to quickly understand the relationship between subject covariates and treatment effects, and to decide whether to trust its estimates.

We describe here a few directions for future work. Firstly, we currently model the effect of confounding covariates under the same grouping structure used for treatment effects. That can be relaxed by, for example, modeling the effect of confounders using a separate model like a Gaussian process while maintaining the same model for treatment effects. In any case, it seems intuitive to give different treatment to confounding features versus treatment effect indicator features. Secondly, the simulated annealing model search can be made more efficient, perhaps using branch and bound techniques to avoid transitions that cannot lead to increased evidence. Lastly, to lessen the upwardly "biased" monotonic priors on subgroup treatment effects explained in the previous section, we may wish to let the prior of each $\delta^{(l)}$ come from a mixture of a point mass at 0 and a positive uniform distribution.

References

Athey, S., and Imbens, G. 2015. Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*.

Beygelzimer, A., and Langford, J. 2009. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 129–138. ACM.

Bishop, C. M., and Svenskn, M. 2002. Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 57–64. Morgan Kaufmann Publishers Inc.

Borgelt, C. 2005. An implementation of the fp-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, 1–5. ACM.

Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9(1):247–274.

Cai, T.; Tian, L.; Wong, P. H.; and Wei, L. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12(2):270–282.

Foster, J. C.; Taylor, J. M.; and Ruberg, S. J. 2011. Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30(24):2867–2880.

Gelman, A., et al. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis* 1(3):515–534.

Imai, K., and Ratkovic, M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Kodratoff, Y. 1994. The comprehensibility manifesto. *KDD Nugget Newsletter* 94(9).

Martens, D.; Vanthienen, J.; Verbeke, W.; and Baesens, B. 2011. Performance of classification models from a user perspective. *Decision Support Systems* 51(4):782–793.

Pazzani, M. J. 2000. Knowledge discovery from data? *Intelligent systems and their applications, IEEE* 15(2):10–12.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics* 34–58.

Sekhon, J. S. 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*.

Su, X., and Nickerson, D. M. 2009. Subgroup Analysis via Recursive Partitioning. 10:141–158.

Su, X.; Kang, J.; and Levine, R. A. 2012. Facilitating Score and Causal Inference Trees for Large Observational Studies. 13:2955–2994.

Su, X.; Wang, M.; and Fan, J. 2004. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics* 13(3):586–598.

Sun, W.; Wang, P.; Yin, D.; Yang, J.; and Chang, Y. 2014. Causal Inference via Sparse Additive Models with Application to Online Advertising.

Taddy, M.; Gardner, M.; Chen, L.; and Draper, D. 2014. Heterogeneous treatment effects in digital experimentation. *arXiv preprint arXiv:1412.8563*.

US Census. 1995. Us census current population survey. Available at http://www.census.gov/cps/.

Wilhelm, S. 2012. Moments calculation for the doubly truncated multivariate normal density. *arXiv preprint arXiv:1206.5387*.

Zeileis, A.; Hothorn, T.; and Hornik, K. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17(2):492–514.