**Dúvidas abordagem da pesquisa:**

Observações:

1. O número de backbones utilizados não mostrou influenciar em nada o resultado das variáveis “desconhecidas”. Concluo que isso seja porque no experimento as variáveis são geradas completamente independente umas das outras, ou seja, mesmo que eu escolha 10 backbones, se as outras 20 variáveis são independentes dessas 10, elas não irão alterar o resultado.
2. O experimento que está sendo feito agora o que consegue-se identificar é qual variável tem uma distribuição mais normal que outra. Entretanto consigo imaginar diversos cenários onde há variáveis com distribuição normal que não sejam importante para o resultado final que se deseja prever.

Exemplo Artificial, pouco provável, porém possível:

Se eu criar um banco de dados artificial, onde a classe é definida por -> f(x1, x2) = sign(x1 – 3x22). E mandar um banco de dados com x1, x2, x3, x4 e f(x1, x2), sendo x1, x2 uniformemente distribuídas e x3 e x4 normalmente distribuídas. Nessa situação nosso algoritmo irá informar que x3 e x4 são backbones quando na verdade nada tem haver com o resultado final.

O que quero tentar explicitar é que penso que, a não ser que os atributos sejam de certa forma dependentes um dos outros (o que nos impede de descobrir backbones independentes) teremos que utilizar aprendizado supervisionado para descobrir os outros backbones, exatamente por causa da formula:

P(x | algo) != P(x).

Para que ela seja verdadeira “algo” tem que alterar a probabilidade de x, e utilizando atributos gerados independentemente a única forma que penso em atingir isso é pegando a classe de “algo” (utilizando aprendizado supervisionado).

PS: tentando explicar minha observação quanto as distâncias.

Acho que não teremos problemas de comprar distâncias com dimensões diferentes, pois, por enquanto elas trabalham em partes distintas do algoritmo.

Exemplo: Se eu utilizar 4 backbones utilizarei a distância no R4 para achar os *n* pontos mais próximos desses backbones. Partir do momento que eu peguei o ponto médio desses *n* pontos, eu comparo as outras variáveis todas em R1. Ou seja o R4 é só para achar os pontos mais próximos, as comparações em si, são feitas no R1.

Mas isso acho que é menos primordial do que as questões acima.