COWS and Their Hybrids

Customized Orthogonal Weights

Larry Wasserman
larry@cmu.edu

Joint work with: Mikael Kuusela and Chad Schafer

sWeights and COWS

sWeights and COWS

- sWeights is an algorithm derived by particle physicists for separating signal from background (Barlow 1987, Pivk and Le Diberder 2005).

sWeights and COWS

- sWeights is an algorithm derived by particle physicists for separating signal from background (Barlow 1987, Pivk and Le Diberder 2005).
- COWs (Customized Orthogonal Weights) is a generalization (Dembinski, Kenzie, Langenbruch and Schmelling 2022).

sWeights and COWS

- sWeights is an algorithm derived by particle physicists for separating signal from background (Barlow 1987, Pivk and Le Diberder 2005).
- COWs (Customized Orthogonal Weights) is a generalization (Dembinski, Kenzie, Langenbruch and Schmelling 2022).
- CS was asked to discuss COWs at a PhyStat meeting.

sWeights and COWS

- sWeights is an algorithm derived by particle physicists for separating signal from background (Barlow 1987, Pivk and Le Diberder 2005).
- COWs (Customized Orthogonal Weights) is a generalization (Dembinski, Kenzie, Langenbruch and Schmelling 2022).
- CS was asked to discuss COWs at a PhyStat meeting.
- This talk explains the method, our statistical deconstruction and some extensions (hybrids).

Setting

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)
  Secondary variables (control) T (decay time)

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)
  Secondary variables (control) T (decay time)
- Model for M

$$p(m) = \underbrace{\pi s(m)}_{\textit{signal}} + \underbrace{(1-\pi)b(m)}_{\textit{background}}$$

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)
  Secondary variables (control) T (decay time)
- Model for M

$$p(m) = \underbrace{\pi s(m)}_{signal} + \underbrace{(1-\pi) b(m)}_{background}$$

- In some cases $s(m)$ is known, or can be simulated.

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)
  Secondary variables (control) T (decay time)
- Model for M

$$p(m) = \underbrace{\pi s(m)}_{signal} + \underbrace{(1-\pi)b(m)}_{background}$$

- In some cases $s(m)$ is known, or can be simulated.
- Or, at least, M can be modeled:

$$p(m) = \pi s(m;\theta) + (1-\pi)b(m;\gamma)$$

Setting

- Two variables:
  Primary variable (discriminant) M (invariant mass)
  Secondary variables (control) T (decay time)
- Model for M

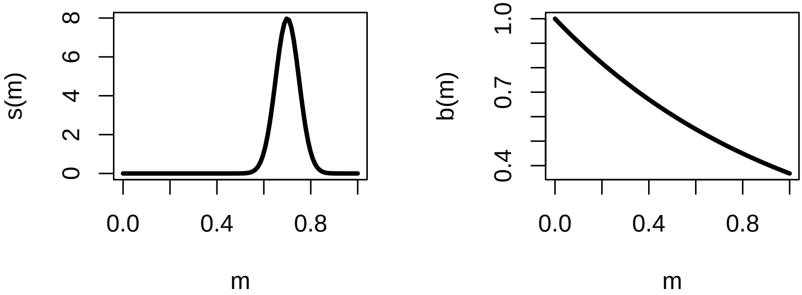
$$p(m) = \underbrace{\pi s(m)}_{signal} + \underbrace{(1-\pi)b(m)}_{background}$$

- In some cases $s(m)$ is known, or can be simulated.
- Or, at least, M can be modeled:

$$p(m) = \pi s(m;\theta) + (1-\pi)b(m;\gamma)$$

- We want to extract the signal component for T.

Signal and Background



sWeights

sWeights

- the sWeights models assumes

$$M \perp\!\!\!\perp T \mid \mathrm{signal} = 1, \qquad M \perp\!\!\!\perp T \mid \mathrm{signal} = 0.$$

sWeights

- the sWeights models assumes

$$M \perp\!\!\!\perp T \mid \text{signal} = 1, \qquad M \perp\!\!\!\perp T \mid \text{signal} = 0.$$

- Hence

$$p(m, t) = \underbrace{\pi g_1(m; \theta) h_1(t)}_{\textit{signal}} + \underbrace{(1 - \pi) g_2(m; \gamma) h_2(t)}_{\textit{background}}$$

sWeights

- the sWeights models assumes

$$M \perp\!\!\!\perp T \mid \text{signal} = 1, \qquad M \perp\!\!\!\perp T \mid \text{signal} = 0.$$

- Hence

$$p(m,t) = \underbrace{\pi g_1(m;\theta) h_1(t)}_{\textit{signal}} + \underbrace{(1-\pi) g_2(m;\gamma) h_2(t)}_{\textit{background}}$$

- This is a semiparametric model: parameters are:

$$\pi, \theta, \gamma, h_1(\cdot), h_2(\cdot)$$

sWeights

- the sWeights models assumes

$$M \perp\!\!\!\perp T \mid \text{signal} = 1, \qquad M \perp\!\!\!\perp T \mid \text{signal} = 0.$$

- Hence

$$p(m,t) = \underbrace{\pi g_1(m;\theta) h_1(t)}_{\textit{signal}} + \underbrace{(1-\pi) g_2(m;\gamma) h_2(t)}_{\textit{background}}$$

- This is a semiparametric model: parameters are:

$$\pi, \theta, \gamma, h_1(\cdot), h_2(\cdot)$$

- The goal is to recover h_1 and h_2.

Identifiability?

Identifiability?

- Recall:

$$p(m, t) = \pi g_1(m)h_1(t) + (1 - \pi)g_2(m)h_2(t).$$

Identifiability?

- Recall:
$$p(m, t) = \pi g_1(m)h_1(t) + (1 - \pi)g_2(m)h_2(t).$$
- This model is identified since we have a model for g_1 and g_2.

Identifiability?

- Recall:
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t).$$
- This model is identified since we have a model for g_1 and g_2.
- Without that marginal model, it is not identified.

Identifiability?

- Recall:
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t).$$
- This model is identified since we have a model for g_1 and g_2.
- Without that marginal model, it is not identified.
- But
$$p(m,t,u) = \pi g_1(m)h_1(t)f_1(u) + (1-\pi)g_2(m)h_2(t)f_2(u)$$
is identified!

Identifiability?

- Recall:
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t).$$
- This model is identified since we have a model for g_1 and g_2.
- Without that marginal model, it is not identified.
- But
$$p(m,t,u) = \pi g_1(m)h_1(t)f_1(u) + (1-\pi)g_2(m)h_2(t)f_2(u)$$
is identified!
- As long as the number of variables $d \geq 3$, the model is nonparametrically identified.

Identifiability?

- Recall:
$$p(m, t) = \pi g_1(m)h_1(t) + (1 - \pi)g_2(m)h_2(t).$$
- This model is identified since we have a model for g_1 and g_2.
- Without that marginal model, it is not identified.
- But
$$p(m, t, u) = \pi g_1(m)h_1(t)f_1(u) + (1 - \pi)g_2(m)h_2(t)f_2(u)$$
is identified!
- As long as the number of variables $d \geq 3$, the model is nonparametrically identified.
- We return to this point later.

sWeights

sWeights

- Choose functions $w_1(m)$ and $w_2(m)$ such that

$$\int w_1(m)g_1(m)dm = 1, \qquad \int w_1(m)g_2(m)dm = 0$$

$$\int w_2(m)g_1(m)dm = 0, \qquad \int w_2(m)g_2(m)dm = 1.$$

sWeights

- Choose functions $w_1(m)$ and $w_2(m)$ such that

$$\int w_1(m)g_1(m)dm = 1, \qquad \int w_1(m)g_2(m)dm = 0$$

$$\int w_2(m)g_1(m)dm = 0, \qquad \int w_2(m)g_2(m)dm = 1.$$

- Then

$$h_1(t) = \mathbb{E}\left[\frac{w_1(M)I(T=t)}{\pi}\right]$$

$$h_2(t) = \mathbb{E}\left[\frac{w_2(M)I(T=t)}{1-\pi}\right].$$

sWeights

- Choose functions $w_1(m)$ and $w_2(m)$ such that

$$\int w_1(m)g_1(m)dm = 1, \qquad \int w_1(m)g_2(m)dm = 0$$

$$\int w_2(m)g_1(m)dm = 0, \qquad \int w_2(m)g_2(m)dm = 1.$$

- Then

$$h_1(t) = \mathbb{E}\left[\frac{w_1(M)I(T=t)}{\pi}\right]$$

$$h_2(t) = \mathbb{E}\left[\frac{w_2(M)I(T=t)}{1-\pi}\right].$$

- Estimators:

$$\widehat{h}_1(t) = \frac{1}{n\widehat{\pi}}\sum_i \widehat{w}_1(M_i)K_h(T_i - t)$$

$$\widehat{h}_2(t) = \frac{1}{n(1-\widehat{\pi})}\sum_i \widehat{w}_2(M_i)K_h(T_i - t).$$

sWeights

- Choose functions $w_1(m)$ and $w_2(m)$ such that

$$\int w_1(m)g_1(m)dm = 1, \qquad \int w_1(m)g_2(m)dm = 0$$
$$\int w_2(m)g_1(m)dm = 0, \qquad \int w_2(m)g_2(m)dm = 1.$$

- Then

$$h_1(t) = \mathbb{E}\left[\frac{w_1(M)I(T=t)}{\pi}\right]$$
$$h_2(t) = \mathbb{E}\left[\frac{w_2(M)I(T=t)}{1-\pi}\right].$$

- Estimators:

$$\widehat{h}_1(t) = \frac{1}{n\widehat{\pi}}\sum_i \widehat{w}_1(M_i)K_h(T_i - t)$$
$$\widehat{h}_2(t) = \frac{1}{n(1-\widehat{\pi})}\sum_i \widehat{w}_2(M_i)K_h(T_i - t).$$

- Under standard conditions, $||\widehat{h}_1 - h_1|| \xrightarrow{P} 0$ and $||\widehat{h}_2 - h_2|| \xrightarrow{P} 0$.

Constructing Weight Functions

Constructing Weight Functions

- Choose ϕ_1, ϕ_2. (Can choose $\phi_1 = g_1$ and $\phi_2 = g_2$.)

Constructing Weight Functions

- Choose ϕ_1, ϕ_2. (Can choose $\phi_1 = g_1$ and $\phi_2 = g_2$.)
- Let

$$w_1(m) = B_{11}\phi_1(m) + B_{12}\phi_2(m)$$
$$w_2(m) = B_{21}\phi_1(m) + B_{22}\phi_2(m)$$

where $B = A^{-1}$ and

$$A_{ij} = \int g_i(m)\phi_j(m)dm.$$

Constructing Weight Functions

- Choose ϕ_1, ϕ_2. (Can choose $\phi_1 = g_1$ and $\phi_2 = g_2$.)
- Let

$$w_1(m) = B_{11}\phi_1(m) + B_{12}\phi_2(m)$$
$$w_2(m) = B_{21}\phi_1(m) + B_{22}\phi_2(m)$$

  where $B = A^{-1}$ and

$$A_{ij} = \int g_i(m)\phi_j(m)dm.$$

- The optimal weight functions depend on h_1 and h_2. (Can be done iteratively).

What Not To Do

What Not To Do

- It is tempting to compute weights

$$\widetilde{w}(M) = P(S = 1 | M = m) = \frac{\pi g_1(m)}{\pi g_1(m) + (1 - \pi) g_2(m)}$$

What Not To Do

- It is tempting to compute weights

$$\widetilde{w}(M) = P(S = 1|M = m) = \frac{\pi g_1(m)}{\pi g_1(m) + (1 - \pi) g_2(m)}$$

- Then use these weights to estimate $h_i(t)$.

What Not To Do

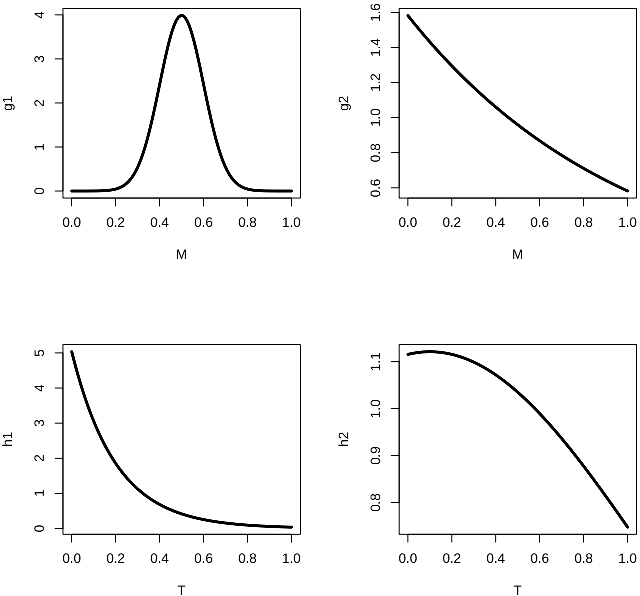
- It is tempting to compute weights

$$\widetilde{w}(M) = P(S = 1|M = m) = \frac{\pi g_1(m)}{\pi g_1(m) + (1-\pi) g_2(m)}$$

- Then use these weights to estimate $h_i(t)$.
- But then

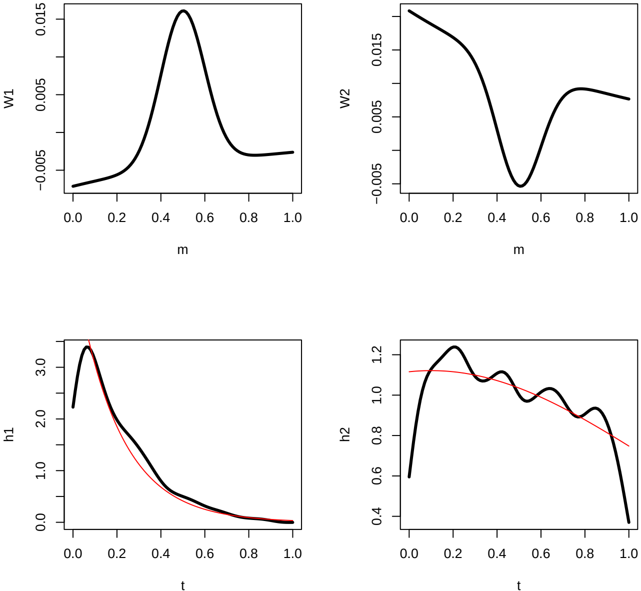
$$\mathbb{E}\left[\frac{\widetilde{w}(M) I(T = t)}{\pi}\right] = h_1(t) \int \frac{\widetilde{w}(m)}{P(S = 1|M = m, T = t)} g_1(m) dm \neq h_1(t).$$

Example



g1
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
M
g2
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0
M
h1
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
T
h2
0.8
0.9
1.0
1.1
0.0
0.2
0.4
0.6
0.8
1.0
T

Example


W1
0.015
0.005
−0.005
0.0
0.2
0.4
0.6
0.8
1.0
m
W2
0.015
0.005
−0.005
m
h1
3.0
2.0
1.0
0.0
t
h2
1.2
1.0
0.8
0.6
0.4
t

COWs

COWs

- Want to relax the conditional independence assumption

COWs

- Want to relax the conditional independence assumption
- The COWs model is

$$p(m,t) = \underbrace{\sum_{j=1}^{s} \pi_j g_j(m) h_j(t)}_{signal} + \underbrace{\sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t)}_{background}.$$

COWs

- Want to relax the conditional independence assumption
- The COWs model is

$$p(m,t) = \underbrace{\sum_{j=1}^{s} \pi_j g_j(m) h_j(t)}_{signal} + \underbrace{\sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t)}_{background}.$$

- We want to estimate

$$h_{signal}(t) = \frac{\sum_{j=1}^{s} \pi_j h_j(t)}{\sum_{j=1}^{s} \pi_j}$$

COWs

- Want to relax the conditional independence assumption
- The COWs model is

$$p(m,t) = \underbrace{\sum_{j=1}^{s} \pi_j g_j(m) h_j(t)}_{signal} + \underbrace{\sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t)}_{background}.$$

- We want to estimate

$$h_{signal}(t) = \frac{\sum_{j=1}^{s} \pi_j h_j(t)}{\sum_{j=1}^{s} \pi_j}$$

- Require marginal model

$$p(m) = \sum_{j=1}^{s} \pi_j g_j(m;\theta_j) + \sum_{j=s+1}^{N} \pi_j g_j(m;\gamma_j).$$

Weights

Weights

- Construct $w_1, \ldots, w_N$ such that

$$\int w_j(m) g_k(m) dm = \delta_{jk}.$$

Weights

- Construct $w_1, \ldots, w_N$ such that

$$\int w_j(m) g_k(m) dm = \delta_{jk}.$$

- Then

$$h_{signal}(t) = \frac{\sum_{j=1}^{s} \mathbb{E}\left[w_j(M) I(T = t)\right]}{\sum_{j=1}^{s} \pi_j}$$

Weights

- Construct $w_1, \dots, w_N$ such that

$$\int w_j(m) g_k(m) dm = \delta_{jk}.$$

- Then

$$h_{signal}(t) = \frac{\sum_{j=1}^{s} \mathbb{E}\left[w_j(M) I(T = t)\right]}{\sum_{j=1}^{s} \pi_j}$$

- Thus

$$\widehat{h}_{signal}(t) = \frac{1}{n} \sum_i K_h(T_t - t) \widehat{w}(M_i)$$

where

$$\widehat{w}(m) = \frac{\sum_{j=1}^{s} \widehat{w}_j(m)}{\sum_{j=1}^{s} \widehat{\pi}_j}.$$

The Herd

The Herd

- Where does the model come from?

The Herd

- Where does the model come from?
- Suppose we are given $\lambda, b(m), s(m)$ so that

$$p(m) = \lambda s(m) + (1 - \lambda) b(m).$$

The Herd

- Where does the model come from?
- Suppose we are given $\lambda, b(m), s(m)$ so that

$$p(m) = \lambda s(m) + (1 - \lambda)b(m).$$

- Find $\{\beta_1, \ldots, \beta_N\}$, $\mathcal{G}_s = \{g_1, \ldots, g_s\}$ and $\mathcal{G}_b = \{g_{s+1}, \ldots, g_N\}$ such that

$$s(m) = \sum_{j=1}^{s} \beta_j g_j(m)$$

$$b(m) = \sum_{j=s+1}^{N} \beta_j g_j(m)$$

The Herd

- Where does the model come from?
- Suppose we are given $\lambda, b(m), s(m)$ so that

$$p(m) = \lambda s(m) + (1 - \lambda) b(m).$$

- Find $\{\beta_1, \ldots, \beta_N\}$, $\mathcal{G}_s = \{g_1, \ldots, g_s\}$ and $\mathcal{G}_b = \{g_{s+1}, \ldots, g_N\}$ such that

$$s(m) = \sum_{j=1}^{s} \beta_j g_j(m)$$

$$b(m) = \sum_{j=s+1}^{N} \beta_j g_j(m)$$

- Require that $\mathcal{G}_s$ and $\mathcal{G}_b$ are linearly independent.

The Herd

The Herd

- This defines a COWs model

$$c(m,t) = \sum_{j=1}^{s} \pi_j g_j(m) h_j(t) + \sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t).$$

The Herd

- This defines a COWs model

$$c(m,t) = \sum_{j=1}^{s} \pi_j g_j(m) h_j(t) + \sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t).$$

- Let $\mathcal{H}$ be the set of all such models such that $c(m,t) = p(m,t)$.

The Herd

- This defines a COWs model

$$c(m,t) = \sum_{j=1}^{s} \pi_j g_j(m) h_j(t) + \sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t).$$

- Let $\mathcal{H}$ be the set of all such models such that $c(m,t) = p(m,t)$.
- Open Questions:
  Is the signal for T $\psi(t) \propto \sum_{j=1}^{s} \pi_j h_j(t)$ identified?

The Herd

- This defines a COWs model

$$c(m,t) = \sum_{j=1}^{s} \pi_j g_j(m) h_j(t) + \sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t).$$

- Let $\mathcal{H}$ be the set of all such models such that $c(m,t) = p(m,t)$.
- Open Questions:
  Is the signal for T $\psi(t) \propto \sum_{j=1}^{s} \pi_j h_j(t)$ identified?
- That is, does every COWs model $\mathcal{H}$ yields the same $\psi(t)$?

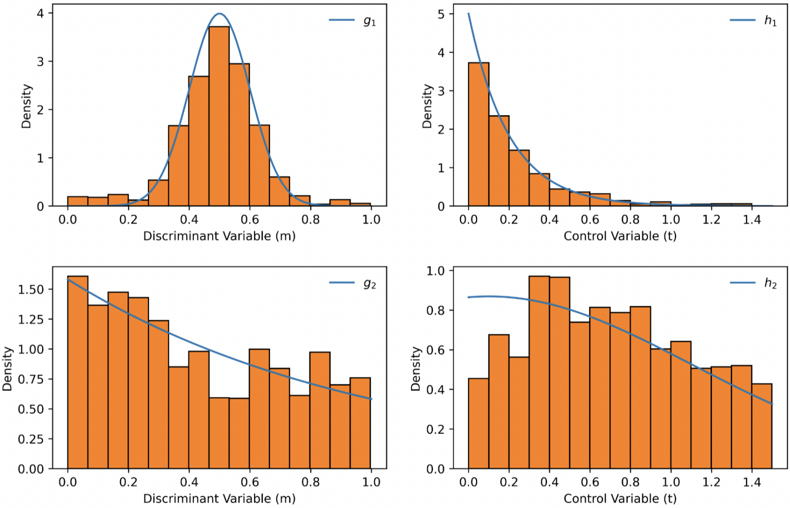
The Herd

- This defines a COWs model

$$c(m,t) = \sum_{j=1}^{s} \pi_j g_j(m) h_j(t) + \sum_{j=s+1}^{N} \pi_j g_j(m) h_j(t).$$

- Let $\mathcal{H}$ be the set of all such models such that $c(m,t) = p(m,t)$.
- Open Questions:
  Is the signal for T $\psi(t) \propto \sum_{j=1}^{s} \pi_j h_j(t)$ identified?
- That is, does every COWs model $\mathcal{H}$ yields the same $\psi(t)$?
- How do we construct the herd?

Example



g_1
Density
Discriminant Variable (m)
h_1
Density
Control Variable (t)
g_2
Density
Discriminant Variable (m)
h_2
Density
Control Variable (t)

Hybrids

Weakening the Assumptions

Weakening the Assumptions

- Recall
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t)$$

Weakening the Assumptions

- Recall
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t)$$
- What if g_1 and g_2 are completely unknown?

Weakening the Assumptions

- Recall
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t)$$
- What if g_1 and g_2 are completely unknown?
- Then h_1 is not identified.

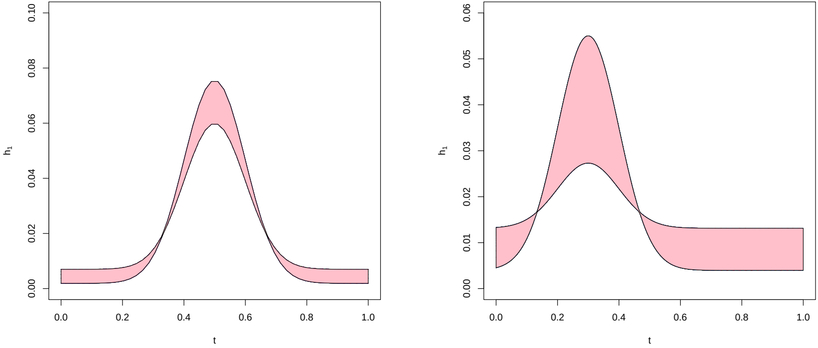
Weakening the Assumptions

- Recall
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t)$$
- What if g_1 and g_2 are completely unknown?
- Then h_1 is not identified.
- But, due to results in Hall and Zhou (2003), there is an explicit one parameter family of possible functions h_1.

Weakening the Assumptions

- Recall
$$p(m,t) = \pi g_1(m)h_1(t) + (1-\pi)g_2(m)h_2(t)$$
- What if g_1 and g_2 are completely unknown?
- Then h_1 is not identified.
- But, due to results in Hall and Zhou (2003), there is an explicit one parameter family of possible functions h_1.
- From this, we can find bounds on $h_1(t)$.

Two Examples



COWs in Higher Dimensions

COWs in Higher Dimensions

- More generally, consider $X_1, \ldots, X_d$ with $d \geq 3$ and k groups.

COWs in Higher Dimensions

- More generally, consider $X_1, \ldots, X_d$ with $d \geq 3$ and k groups.
- So
$$p(x_1, \ldots, x_d) = \sum_{j=1}^{k} \pi_j \prod_{r=1}^{d} f_{jr}(x_r).$$

COWs in Higher Dimensions

- More generally, consider $X_1, \ldots, X_d$ with $d \geq 3$ and k groups.
- So
$$p(x_1, \ldots, x_d) = \sum_{j=1}^{k} \pi_j \prod_{r=1}^{d} f_{jr}(x_r).$$
- Parameters are $\pi = (\pi_1, \ldots, \pi_d)$ and
$$F = \{f_{jr} :\ 1 \leq j \leq k,\ 1 \leq r \leq d\}.$$

COWs in Higher Dimensions

- More generally, consider $X_1, \ldots, X_d$ with $d \geq 3$ and k groups.
- So
$$p(x_1, \ldots, x_d) = \sum_{j=1}^{k} \pi_j \prod_{r=1}^{d} f_{jr}(x_r).$$
- Parameters are $\pi = (\pi_1, \ldots, \pi_d)$ and
$$F = \{f_{jr} :\ 1 \leq j \leq k,\ 1 \leq r \leq d\}.$$
- This model is nonparametrically identified. (Hall and Zhou 2003, Allman et al 2009, Levine, Hunter, Chaveau 2011, Bonhomme et al 2016, Tang et al 2025, Zheng and Wu 2020, Chaumaray and Marbac 2024).

COWs in Higher Dimensions

- More generally, consider $X_1, \ldots, X_d$ with $d \geq 3$ and k groups.
- So
$$p(x_1, \ldots, x_d) = \sum_{j=1}^{k} \pi_j \prod_{r=1}^{d} f_{jr}(x_r).$$
- Parameters are $\pi = (\pi_1, \ldots, \pi_d)$ and
$$F = \{f_{jr} :\ 1 \leq j \leq k,\ 1 \leq r \leq d\}.$$
- This model is nonparametrically identified. (Hall and Zhou 2003, Allman et al 2009, Levine, Hunter, Chaveau 2011, Bonhomme et al 2016, Tang et al 2025, Zheng and Wu 2020, Chaumaray and Marbac 2024).
- COWs offers a different (easier) way to fit this model.

Existing Methods: EM

Existing Methods: EM

- (Levine, Hunter et al) Maximize the smoothed log-likelihood

$$\ell(g) = \int g(x) \log \sum_j \pi_j \mathcal{N}(f_j)(x) dx$$

where

$$\mathcal{N}(f)(x) = e^{\int K_h(u-x) \log f(u) du}.$$

Existing Methods: EM

- (Levine, Hunter et al) Maximize the smoothed log-likelihood

$$\ell(g) = \int g(x) \log \sum_j \pi_j \mathcal{N}(f_j)(x) dx$$

  where

$$\mathcal{N}(f)(x) = e^{\int K_h(u-x) \log f(u) du}.$$

- Then iterate:

$$w_{ij} = \frac{\pi_j \prod_{r=1}^d \mathcal{N}(f_{jr})(X_{ir})}{\sum_j \pi_j \prod_{r=1}^d \mathcal{N}(f_{jr})(X_{ir})}$$

$$\pi_j = \frac{1}{n} \sum_i w_{ij}$$

$$f_{jr}(u) = \frac{1}{nh\pi_j} \sum_i \sum_r w_{ij} K_h(u - X_{ir}).$$

Other Methods

Other Methods

- Bonhomme et al (2016).

Other Methods

- Bonhomme et al (2016).
- Consider the discrete case:

$$p = \sum_j \pi_j (f_1 \otimes \cdots \otimes f_d).$$

Other Methods

- Bonhomme et al (2016).
- Consider the discrete case:

$$p = \sum_j \pi_j (f_1 \otimes \cdots \otimes f_d).$$

- They show that slices of p and marginals yield matrices that can be jointly diagonalized in a common basis.

Other Methods

- Bonhomme et al (2016).
- Consider the discrete case:

$$p = \sum_j \pi_j (f_1 \otimes \cdots \otimes f_d).$$

- They show that slices of p and marginals yield matrices that can be jointly diagonalized in a common basis.
- The common diagonal matrix contains the component densities.

Other Methods

- Bonhomme et al (2016).
- Consider the discrete case:

$$p = \sum_j \pi_j (f_1 \otimes \cdots \otimes f_d).$$

- They show that slices of p and marginals yield matrices that can be jointly diagonalized in a common basis.
- The common diagonal matrix contains the component densities.
- Apply specialized matrix optimization to perform simultaneous diagonalization

Other Methods

- Bonhomme et al (2016).
- Consider the discrete case:

$$p = \sum_j \pi_j (f_1 \otimes \cdots \otimes f_d).$$

- They show that slices of p and marginals yield matrices that can be jointly diagonalized in a common basis.
- The common diagonal matrix contains the component densities.
- Apply specialized matrix optimization to perform simultaneous diagonalization
- Requires specialized optimization methods

Other Methods

Other Methods

- Zheng and Wu (2020). Partition sample space into pieces.

Other Methods

- Zheng and Wu (2020). Partition sample space into pieces.
  - Conditional densities given partition provide basis functions.

Other Methods

- Zheng and Wu (2020). Partition sample space into pieces.
  - Conditional densities given partition provide basis functions.
  - Estimate the component densities by minimizing a tensor equation using alternating least squares.

Other Methods

- Zheng and Wu (2020). Partition sample space into pieces.
  - Conditional densities given partition provide basis functions.
  - Estimate the component densities by minimizing a tensor equation using alternating least squares.
- Chaumaray and Marbac (2024). Sieve approach. Bin the data. Get mle. Let number of bins increase.

COWs

COWs

- As before, specify model for one margin X_ℓ.

COWs

- As before, specify model for one margin X_ℓ.
- This defines weight functions $\mathcal{W}_\ell = \{w_1, \ldots, w_k\}$ such that

$$\int f_{j\ell}(x_\ell) w_r(x_\ell) dx_\ell = \delta_{jr}$$

COWs

- As before, specify model for one margin X_ℓ.
- This defines weight functions $\mathcal{W}_\ell = \{w_1, \ldots, w_k\}$ such that

$$\int f_{j\ell}(x_\ell) w_r(x_\ell) dx_\ell = \delta_{jr}$$

- Thus

$$f_{js}(x_s) = \mathbb{E}\left[\frac{w_j(X_\ell) I(X_s = x_s)}{\pi_j}\right], \quad \widehat{f}_{js}(x_s) = \frac{1}{n}\sum_i \frac{w_j(X_{i\ell}) K_h(X_{is} - x_s)}{\widehat{\pi}_j}.$$

COWs

- As before, specify model for one margin X_ℓ.
- This defines weight functions $\mathcal{W}_\ell = \{w_1, \ldots, w_k\}$ such that

$$\int f_{j\ell}(x_\ell) w_r(x_\ell) dx_\ell = \delta_{jr}$$

- Thus

$$f_{js}(x_s) = \mathbb{E}\left[\frac{w_j(X_\ell) I(X_s = x_s)}{\pi_j}\right], \quad \widehat{f}_{js}(x_s) = \frac{1}{n}\sum_i \frac{w_j(X_{i\ell}) K_h(X_{is} - x_s)}{\widehat{\pi}_j}.$$

- No iteration. No complex optimization.

COWs

- As before, specify model for one margin X_ℓ.
- This defines weight functions $\mathcal{W}_\ell = \{w_1, \ldots, w_k\}$ such that

$$\int f_{j\ell}(x_\ell) w_r(x_\ell) dx_\ell = \delta_{jr}$$

- Thus

$$f_{js}(x_s) = \mathbb{E}\left[\frac{w_j(X_\ell) I(X_s = x_s)}{\pi_j}\right], \quad \widehat{f}_{js}(x_s) = \frac{1}{n}\sum_i \frac{w_j(X_{i\ell}) K_h(X_{is} - x_s)}{\widehat{\pi}_j}.$$

- No iteration. No complex optimization.
- Two hybrids:
  - agnostic COWs
  - robust COWs

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

- Specify first column (starting values).

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

- Specify first column (starting values).
- Get weights. Run COWs. Get matrix F.

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

- Specify first column (starting values).
- Get weights. Run COWs. Get matrix F.
- Extract second column.

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

- Specify first column (starting values).
- Get weights. Run COWs. Get matrix F.
- Extract second column.
- Get weights. Run COWs. Get F.

Agnostic COWs: No Marginal Model

	1	2	$\cdots$	d
1	f_{11}	f_{12}	$\cdots$	f_{1d}
2	f_{21}	f_{22}	$\cdots$	f_{2d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_{k1}	f_{k2}	$\cdots$	f_{kd}

- Specify first column (starting values).
- Get weights. Run COWs. Get matrix F.
- Extract second column.
- Get weights. Run COWs. Get F.
- Extract third column.

Agnostic COWs: No Marginal Model

$$\begin{array}{c|cccc} & 1 & 2 & \cdots & \mathrm{d} \\ \hline 1 & f_{11} & f_{12} & \cdots & f_{1d} \\ 2 & f_{21} & f_{22} & \cdots & f_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{k} & f_{k1} & f_{k2} & \cdots & f_{kd} \end{array}$$

- Specify first column (starting values).
- Get weights. Run COWs. Get matrix F.
- Extract second column.
- Get weights. Run COWs. Get F.
- Extract third column.
- etc

True densities are a fixed point.

Robust COWs

Robust COWs

- Specify a model for each margin:

$$p_r(x_r) = \sum_j \pi_j f(x_r; \theta_r)$$

Robust COWs

- Specify a model for each margin:

$$p_r(x_r) = \sum_j \pi_j f(x_r; \theta_r)$$

- Each marginal model gives weight function w. Run COWs to get estimate $\widehat{F}^r$ and $\widehat{\pi}^r$. This gives

$$\widehat{p}^r = \sum_j \widehat{\pi}_j^r \prod_s \widehat{f}_{js}^r.$$

Robust COWs

- Specify a model for each margin:

$$p_r(x_r) = \sum_j \pi_j f(x_r; \theta_r)$$

- Each marginal model gives weight function w. Run COWs to get estimate $\widehat{F}^r$ and $\widehat{\pi}^r$. This gives

$$\widehat{p}^r = \sum_j \widehat{\pi}_j^r \prod_s \widehat{f}_{js}^r.$$

- Select $\widehat{r}$ to minimize L_2 error:

$$\int (\widehat{p}^r)^2 - \frac{2}{n} \sum_i \widehat{p}^r(X_i).$$

Robust COWs

- Specify a model for each margin:

$$p_r(x_r) = \sum_j \pi_j f(x_r; \theta_r)$$

- Each marginal model gives weight function w. Run COWs to get estimate $\widehat{F}^r$ and $\widehat{\pi}^r$. This gives

$$\widehat{p}^r = \sum_j \widehat{\pi}_j^r \prod_s \widehat{f}_{js}^r.$$

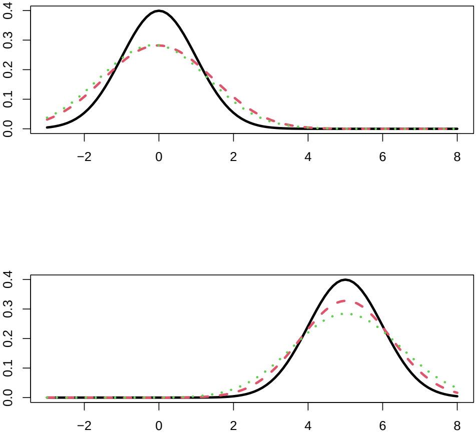
- Select $\widehat{r}$ to minimize L_2 error:

$$\int (\widehat{p}^r)^2 - \frac{2}{n} \sum_i \widehat{p}^r(X_i).$$

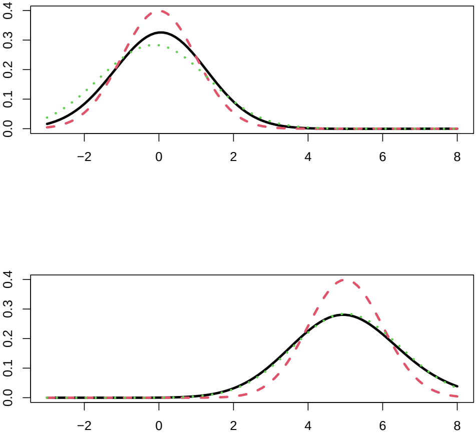
- Multiple robust: If at least one model is correct then $||\widehat{p}^{\widehat{r}} - p|| \xrightarrow{P} 0$ and

$$||\widehat{p}^{\widehat{r}} - p|| = O_P(n^{-1/(2\beta+1)}).$$

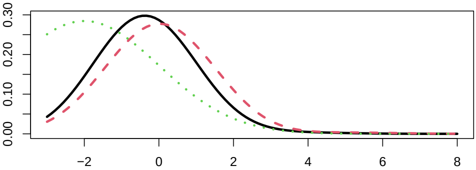
Example: Three Models, Third Misspecified

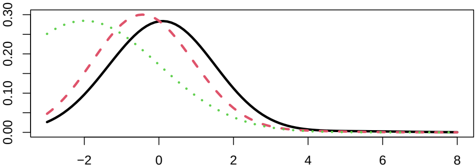


Example: Three Models, Third Misspecified

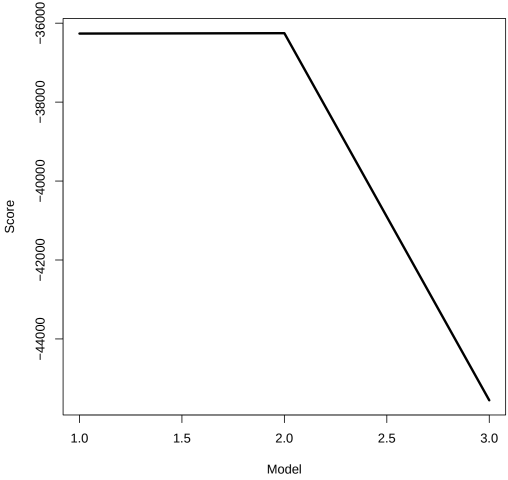


Example: Three Models, Third Misspecified





Example: Three Models, Third Misspecified



Bonus: Alternating Least Squares

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

- Minimize

$$||\widehat{p} - \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})||^2$$

  iteratively for each f_{jr}.

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

- Minimize

$$||\widehat{p} - \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})||^2$$

  iteratively for each f_{jr}.
- Let P_r be the matrix obtained by flattening the array p.

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

- Minimize

$$||\widehat{p} - \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})||^2$$

  iteratively for each f_{jr}.
- Let P_r be the matrix obtained by flattening the array p.
- Let $G_r = [f_{r1}, \ldots, f_{rk}]$ and let $G_{-r} = [W_1, \ldots, W_k]$ where

$$W_j = \mathrm{vec}(f_{j1} \otimes \cdots \otimes f_{j,r-1} \otimes f_{j,r+1} \otimes f_{j,d}).$$

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

- Minimize

$$||\widehat{p} - \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})||^2$$

  iteratively for each f_{jr}.
- Let P_r be the matrix obtained by flattening the array p.
- Let $G_r = [f_{r1}, \ldots, f_{rk}]$ and let $G_{-r} = [W_1, \ldots, W_k]$ where

$$W_j = \mathrm{vec}(f_{j1} \otimes \cdots \otimes f_{j,r-1} \otimes f_{j,r+1} \otimes f_{j,d}).$$

- So

$$P_r = G_r \Pi G_{-r}^T$$

  where $\Pi = \mathrm{diag}(\pi)$.

Bonus: Alternating Least Squares

- Estimators can also be derived using alternating least squares.
- Write p as a mixture of tensors:

$$p = \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})$$

- Minimize

$$||\widehat{p} - \sum_j \pi_j (f_{j1} \otimes \cdots \otimes f_{jd})||^2$$

  iteratively for each f_{jr}.
- Let P_r be the matrix obtained by flattening the array p.
- Let $G_r = [f_{r1}, \ldots, f_{rk}]$ and let $G_{-r} = [W_1, \ldots, W_k]$ where

$$W_j = \mathrm{vec}(f_{j1} \otimes \cdots \otimes f_{j,r-1} \otimes f_{j,r+1} \otimes f_{j,d}).$$

- So

$$P_r = G_r \Pi G_{-r}^T$$

  where $\Pi = \mathrm{diag}(\pi)$.
- Then

$$G_r = P_r (G_{-r}^T G_{-r})^{-1} \Pi^{-1}.$$

Localized Signals

Localized Signals

- Often the signal is localized.

Localized Signals

- Often the signal is localized.
- In this case, there is yet another way to relax conditional independence.

Localized Signals

- Often the signal is localized.
- In this case, there is yet another way to relax conditional independence.
- Write

$$
\begin{aligned}
p(m,t) &= \pi g_1(m) h_1(t|m) + (1-\pi) g_2(m) h_2(t|m) \\
&= \pi s(m,t) + (1-\pi) b(m,t)
\end{aligned}
$$

Localized Signals

- Often the signal is localized.
- In this case, there is yet another way to relax conditional independence.
- Write

$$\begin{aligned} p(m,t) &= \pi g_1(m)h_1(t|m) + (1-\pi)g_2(m)h_2(t|m) \\ &= \pi s(m,t) + (1-\pi)b(m,t) \end{aligned}$$

- The signal $g_1(m)$ is localized to a region $S = [c_1, c_2]$.

Localized Signal

Localized Signal

- For $m \in S^c$ there is no signal (first term is 0) so we can estimate $b(t, m)$ for $t \in S^c$.

Localized Signal

- For $m \in S^c$ there is no signal (first term is 0) so we can estimate $b(t, m)$ for $t \in S^c$.
- And, for $m \in S^c$,

$$b(t, m) = h_2(t|m) g_2(m)$$

so that

$$h_2(t|m) = \frac{b(t, m)}{g_2(m)}.$$

Localized Signal

- For $m \in S^c$ there is no signal (first term is 0) so we can estimate $b(t, m)$ for $t \in S^c$.
- And, for $m \in S^c$,
$$b(t, m) = h_2(t|m) g_2(m)$$
so that
$$h_2(t|m) = \frac{b(t, m)}{g_2(m)}.$$
- We interpolate $h_2(t|m)$ for $c_1 \leq m \leq c_2$ using optimal transport.

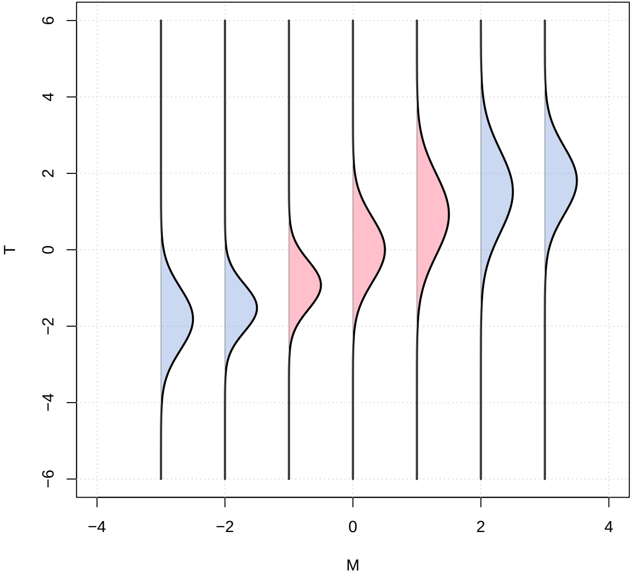
Localized Signal

- For $m \in S^c$ there is no signal (first term is 0) so we can estimate $b(t, m)$ for $t \in S^c$.
- And, for $m \in S^c$,
$$b(t, m) = h_2(t|m)g_2(m)$$
so that
$$h_2(t|m) = \frac{b(t, m)}{g_2(m)}.$$
- We interpolate $h_2(t|m)$ for $c_1 \leq m \leq c_2$ using optimal transport.
- Given $h_2(t|c_1)$ and $h_2(t|c_2)$ there is a shortest path (in Wasserstein distance) $(h_2(t|c) : \ c_1 \leq c \leq c_2)$.

Localized Signal

- For $m \in S^c$ there is no signal (first term is 0) so we can estimate $b(t, m)$ for $t \in S^c$.
- And, for $m \in S^c$,
$$b(t, m) = h_2(t|m) g_2(m)$$
so that
$$h_2(t|m) = \frac{b(t, m)}{g_2(m)}.$$
- We interpolate $h_2(t|m)$ for $c_1 \leq m \leq c_2$ using optimal transport.
- Given $h_2(t|c_1)$ and $h_2(t|c_2)$ there is a shortest path (in Wasserstein distance) $(h_2(t|c) : \; c_1 \leq c \leq c_2)$.
- Then
$$s(m, t) = \frac{p(m, t) - \pi g_2(m) h_2(t|m)}{1 - \pi}.$$

Localized Signal



Semiparametric Copulas

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0, 1]$ with uniform marginals.

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0, 1]$ with uniform marginals.
- Any density $p(x, y)$ can be written

$$p(x, y)p_X(x)p_Y(y)c(F_X(x), F_Y(y)).$$

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0,1]$ with uniform marginals.
- Any density $p(x,y)$ can be written

$$p(x,y)p_X(x)p_Y(y)c(F_X(x),F_Y(y)).$$

- We take

$$p(m,t)=\pi g_1(m)h_1(t)c(G_1(m),H_1(t);\psi_1)+(1-\pi)g_2(m)h_2(t)c(G_2(m),H_2(t);\psi_2).$$

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0, 1]$ with uniform marginals.
- Any density $p(x, y)$ can be written

$$p(x, y)p_X(x)p_Y(y)c(F_X(x), F_Y(y)).$$

- We take

$$p(m, t) = \pi g_1(m)h_1(t)c(G_1(m), H_1(t); \psi_1)+(1-\pi)g_2(m)h_2(t)c(G_2(m), H_2(t); \psi_2).$$

- This allows dependence without having to add more terms.

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0, 1]$ with uniform marginals.
- Any density $p(x, y)$ can be written

$$p(x, y)p_X(x)p_Y(y)c(F_X(x), F_Y(y)).$$

- We take

$$p(m, t) = \pi g_1(m)h_1(t)c(G_1(m), H_1(t); \psi_1) + (1-\pi)g_2(m)h_2(t)c(G_2(m), H_2(t); \psi_2).$$

- This allows dependence without having to add more terms.
- Can be estimated using an EM algorithm.

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0,1]$ with uniform marginals.
- Any density $p(x,y)$ can be written

$$p(x,y)p_X(x)p_Y(y)c(F_X(x),F_Y(y)).$$

- We take

$$p(m,t) = \pi g_1(m)h_1(t)c(G_1(m),H_1(t);\psi_1)+(1-\pi)g_2(m)h_2(t)c(G_2(m),H_2(t);\psi_2).$$

- This allows dependence without having to add more terms.
- Can be estimated using an EM algorithm.
- sWeights provides good starting values of h_1, h_2.

Semiparametric Copulas

- Recall that a copula C with density c is a distribution on $[0, 1]$ with uniform marginals.
- Any density $p(x, y)$ can be written

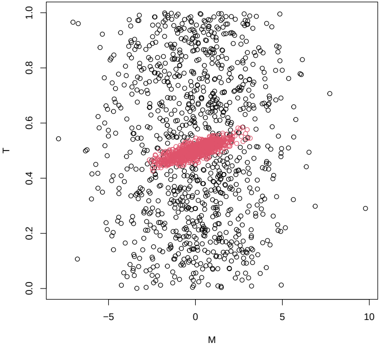
$$p(x, y)p_X(x)p_Y(y)c(F_X(x), F_Y(y)).$$

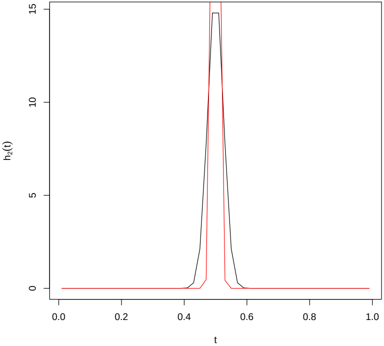
- We take

$$p(m, t) = \pi g_1(m)h_1(t)c(G_1(m), H_1(t); \psi_1) + (1-\pi)g_2(m)h_2(t)c(G_2(m), H_2(t); \psi_2).$$

- This allows dependence without having to add more terms.
- Can be estimated using an EM algorithm.
- sWeights provides good starting values of h_1, h_2.
- EM is especially easy for the Gaussian copula.

Example: Overlapping Signals





Conclusion

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.
- Allows one to leverage information about one marginal.

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.
- Allows one to leverage information about one marginal.
- The extensions are useful more broadly.

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.
- Allows one to leverage information about one marginal.
- The extensions are useful more broadly.
- Physics is a good source of interesting statistical problems:

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.
- Allows one to leverage information about one marginal.
- The extensions are useful more broadly.
- Physics is a good source of interesting statistical problems:
  - Phystat: https://phystat.github.io
  - STAMPS: https://www.cmu.edu/dietrich/statistics-datascience/stamps/

Conclusion

- COWs and the hybrids provide a set of tools for separating signal and background.
- Allows one to leverage information about one marginal.
- The extensions are useful more broadly.
- Physics is a good source of interesting statistical problems:
  - Phystat: https://phystat.github.io
  - STAMPS: https://www.cmu.edu/dietrich/statistics-datascience/stamps/

THE END