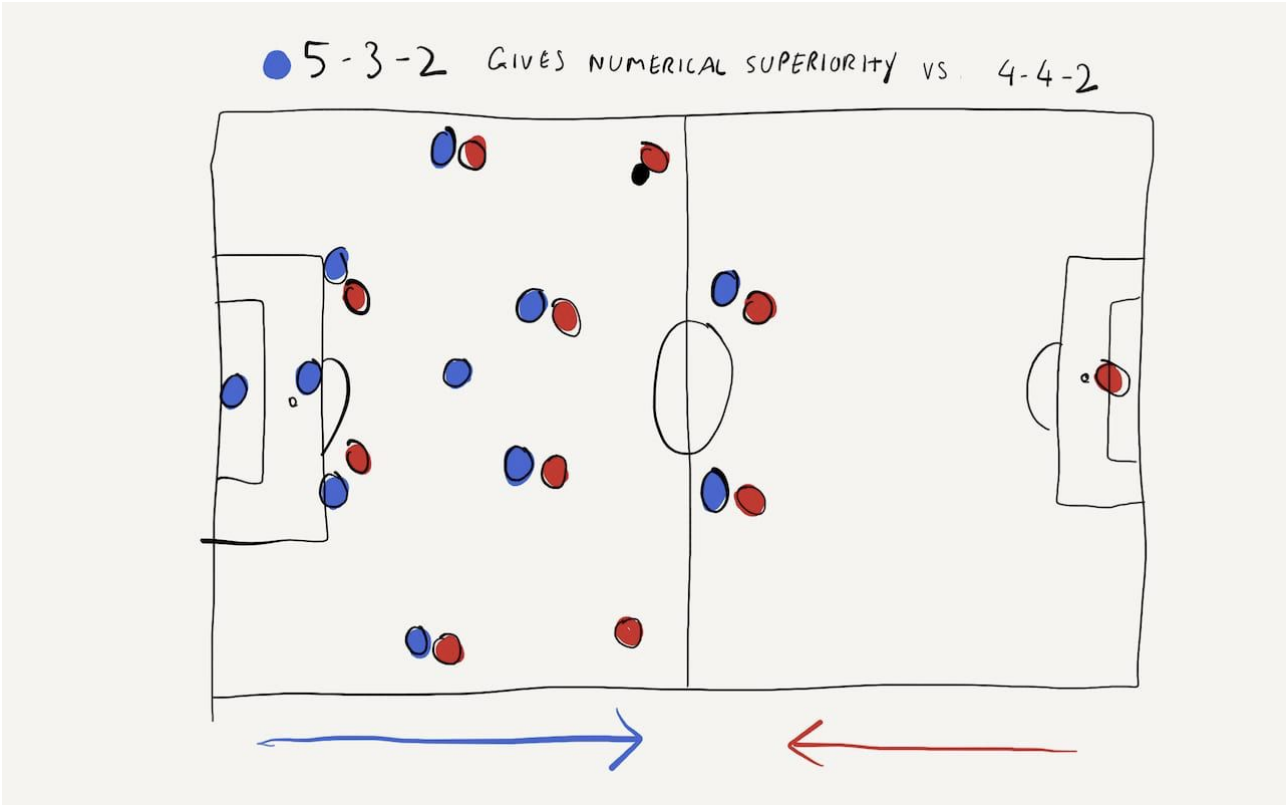
# Identifying soccer formations using broadcast tracking data

Jacob Mortensen - CASSIS 2022


# Background

# Soccer Formations



# Previous Work

Bialkowski et. al. (2014) - Apply minimum entropy data partitioning (MEDP) to assign players to roles and use agglomerative hierarchical clustering to classify roles into formations

Shaw and Glickman (2019) - Average player locations over two minute segments into observations and then perform hierarchical clustering.

# Learning formations from data

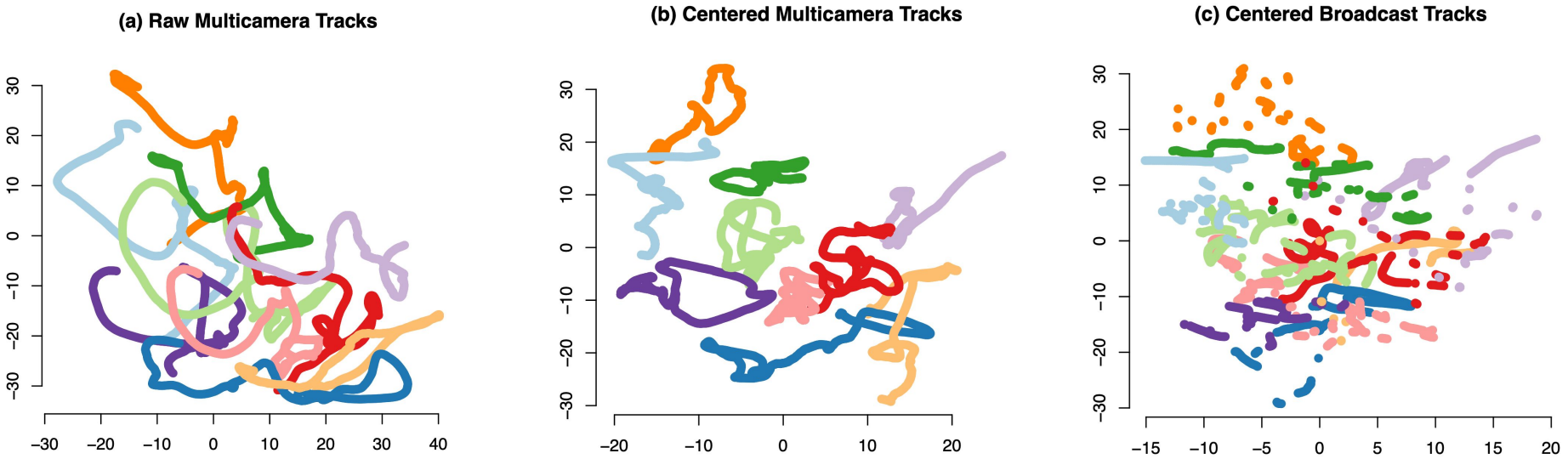
General approach:

1. Aggregate frames to create “observations”
2. Perform some kind of clustering

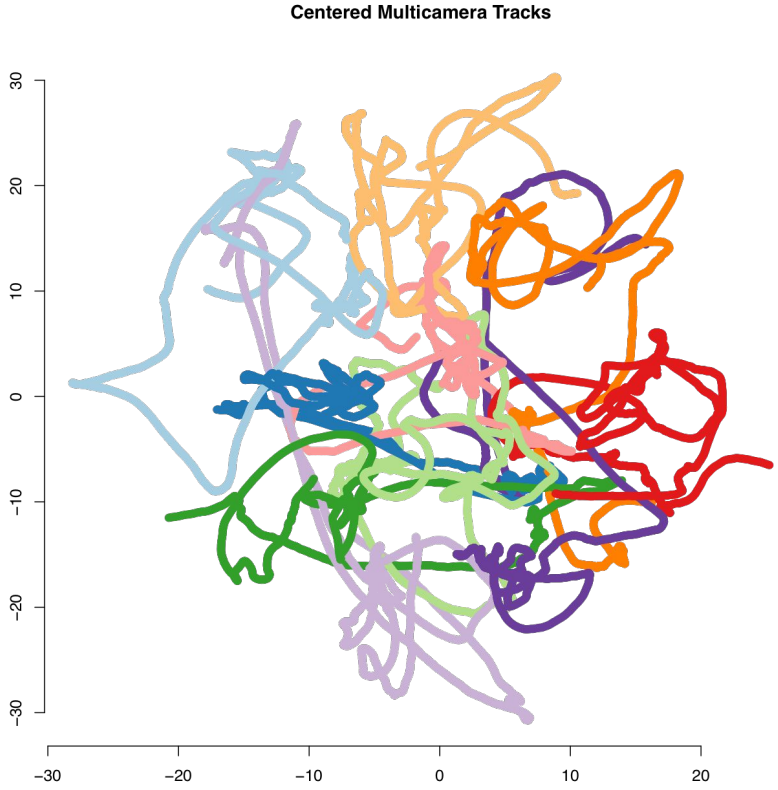
Nuisance factors:

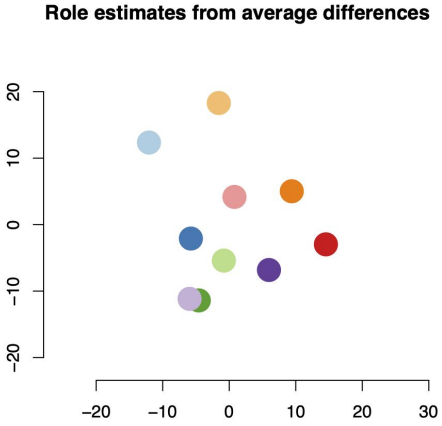
- Frame alignment
  - Broadcast data has the added complication of heavy censoring.
- Player permutation
- Contraction and expansion

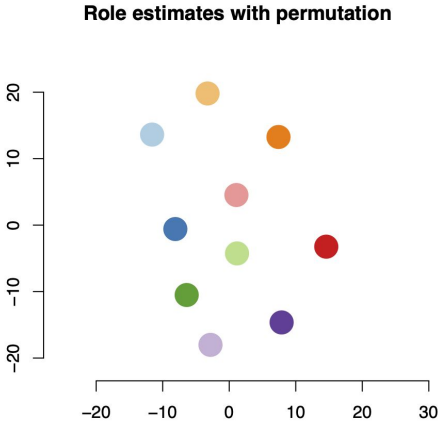
# Frame Alignment



# Permutation



# Data

# Data

- 5 games in a high level European soccer league w/ both optical tracking and broadcast tracking data
- x- and y- coordinates for 11 players on each team, at a frequency of 24 frames per second
- team IDs, player IDs, and possession indicator
- Camera window from broadcast tracking data is used to censor multicamera tracking data, creating a set of “perfect” broadcast tracking data
- **Heavily censored:** the largest percentage of frames where all 10 outfield players were observed for a given team was just 8.48 percent.

# Method

# Atomic Configuration Distance

ACD is a permutation invariant distance measure that calculates the distance between two collections of points by associating a density with each collection of points and calculating the L2 norm between the densities.

$$d^{(acd)}(X,Y) = ||\rho_X(s,\Sigma) - \rho_Y(s,\Sigma)||_2^2$$

$$\rho_X(s,\Sigma) = \frac{1}{N_x}\sum_{i=1}^{N_x}\phi(s;x_i,\Sigma)$$

# Atomic Configuration Distance

$$
\begin{aligned}
d^{(acd)}(X, Y) &= ||\rho_X(s, \Sigma) - \rho_Y(s, \Sigma)||_2^2 \\
&= \int_{s \in \mathbb{R}^D} (\rho_X(s, \Sigma) - \rho_Y(s, \Sigma))^2 ds \\
&= S(\rho_X, \rho_X) + S(\rho_Y, \rho_Y) - 2S(\rho_X, \rho_Y)
\end{aligned}
$$

Miller and Bornn (2020) showed that

$$
S(\rho_X, \rho_Y) = \left(\frac{1}{4\pi}\right)^{D/2} \frac{|\Sigma|^{-1/2}}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \exp\left(-\frac{1}{4}(x_i - y_j)' \Sigma^{-1} (x_i - y_j)\right)
$$

# Atomic Configuration Distance

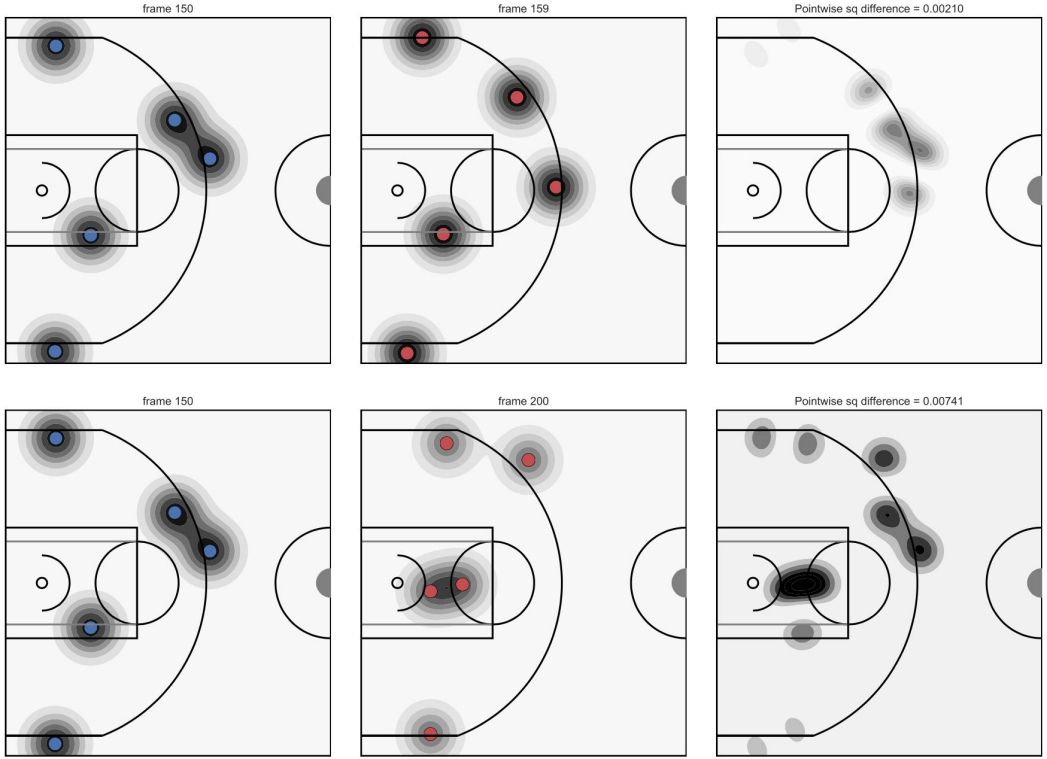



Image taken from Miller and Bornn (2020)

Solves for two nuisance factors:

- Permutation invariant
- Valid if Nx != Ny

# Minimum Entropy Data Partitioning (MEDP)

Let the formation density be $P(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} P_n(\mathbf{x})$ and maximize $\sum_{i=1}^{N} KL(P_n||P)$

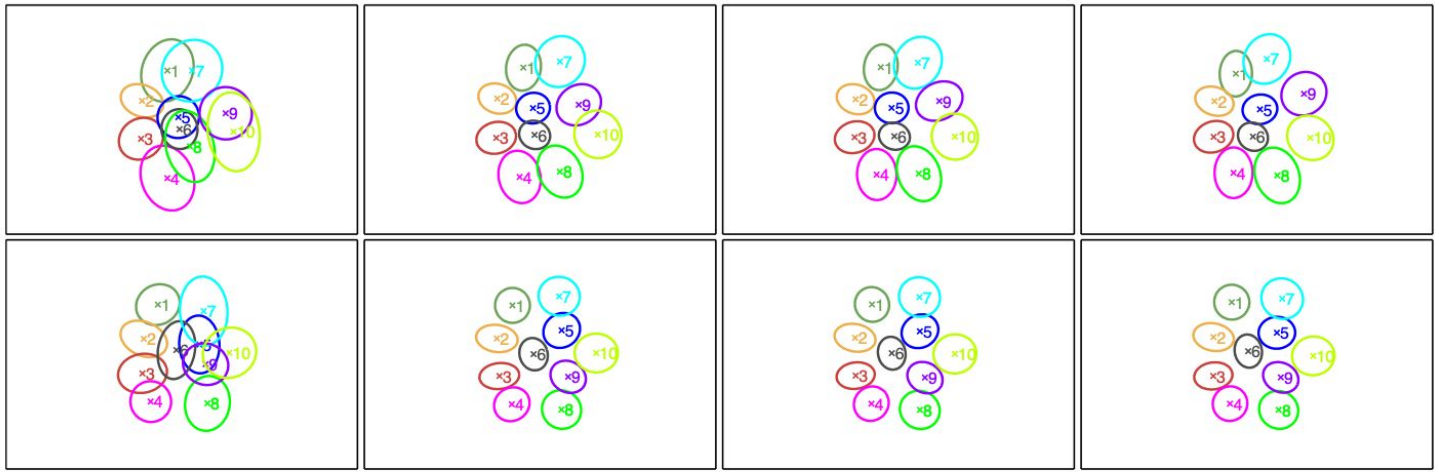



Image taken from Bialkowski et. al. 2014

# Minimum Entropy Data Partitioning (MEDP)

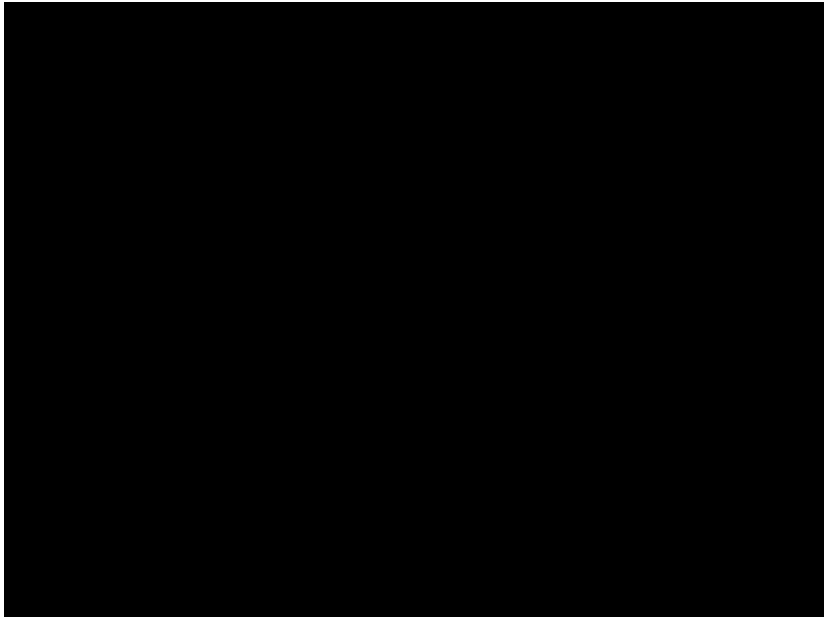
In practice, KL cannot be maximized efficiently, so we follow Bialkowski et. al. (2014) and take an EM-like approach.

1. Initialize algorithm by assuming that players maintain the same role throughout the segment
2. Estimate 2d Gaussian for each role based on centered player locations
3. Calculate cost matrix by evaluating the role densities at each player location and use Hungarian algorithm to permute players with roles to minimize cost
4. Repeat 2-3 until no further reassignments occur

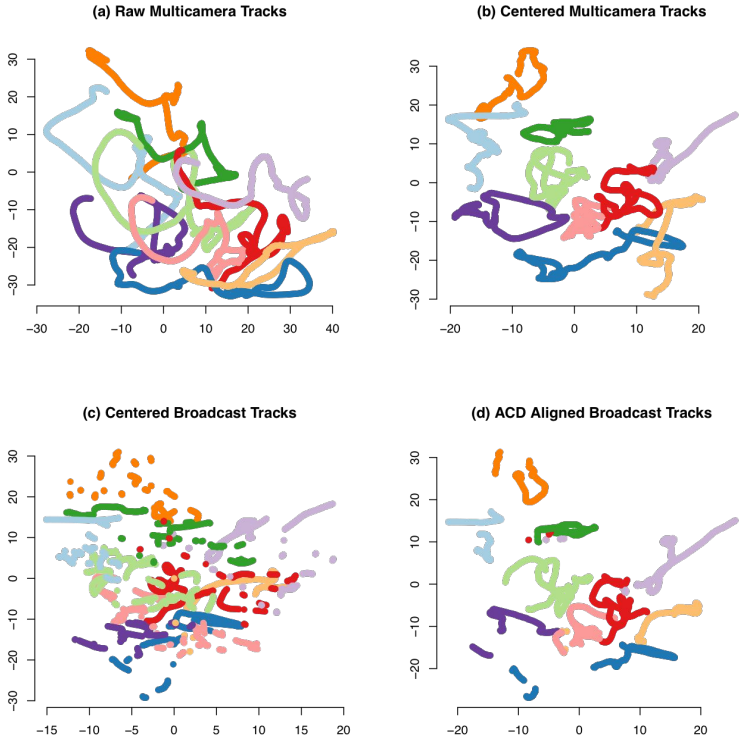
NOTE: We adapt the Hungarian algorithm for use with censored data by setting cost to infinity for missing players. This ensures that observed player locations get assigned to the “cheapest” role density.

# Frame Alignment Algorithm

1. Take a sequence $S_1$, …, $S_T$ where at least 1 player is observed in every frame
2. Find the $\delta_t$ that minimizes the ACD between $S_t + \delta_t$ and $S_{t+1}$ for all t=1,...,T-1, where $\delta_t$ is a matrix with all rows equal to $(\delta_{tx}, \delta_{ty})$
3. Calculate translated frames $S_{\Delta t} = S_t + \sum_{j=t}^{T-1} \delta_j$
4. Use MEDP to calculate role distribution for sequence, and use centroid of role distribution to translate sequence to origin



# Frame Alignment Algorithm



# Aggregate and Scale

- Take similar approach to Shaw and Glickman and aggregate frames over 120 second windows, but we use MEDP rather than average difference approach
  - Average offense and defense separately
  - Exclude possessions < 5 seconds in length
  - Do not average across periods or player substitutions
- Scale data by empirical standard deviation of x- and y-coordinates

# Cluster using Gaussian Mixture Model

We fit a standard GMM using expectation maximization with some small adaptations:

- Hungarian algorithm is used to account for the effects of player permutation
- EM has to be adapted for incomplete observations following Gharamani and Jordan (1994):

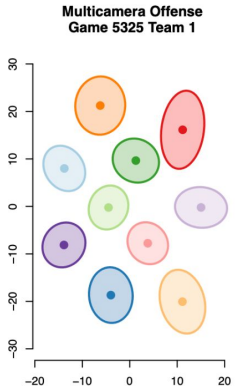
$$\hat{z}_{ik} = P(z_i = k) = \frac{\prod_{p=1}^{2P} \left(\mathcal{N}(x_{ip}|\mu_{kp},\sigma_{kp}^2)\right)^{I(o_{ip}=1)}}{\sum_{\ell=1}^{K} \prod_{p=1}^{2P} \left(\mathcal{N}(x_{ip}|\mu_{\ell p},\sigma_{\ell p}^2)\right)^{I(o_{ip}=1)}}$$

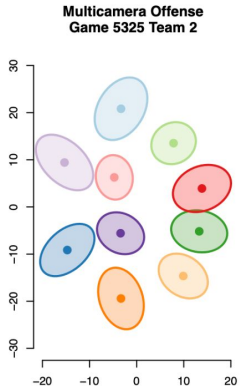
  We replace $I(o_{ip} = 1)$ with the weight $w_{ip} = \sum_{n=1}^{N_i} o_{np}$

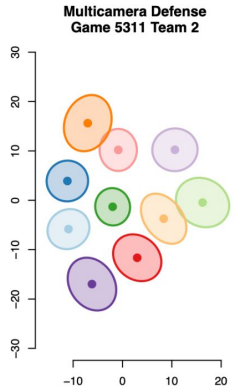
Number of clusters is selected using BIC. Determined to be 8 clusters for defense and 6 clusters for offense.

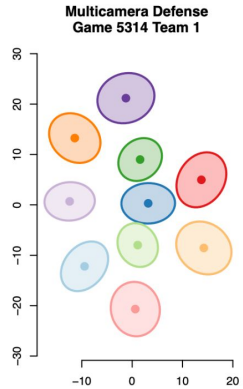
# Results

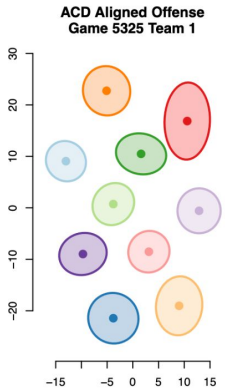
# Results

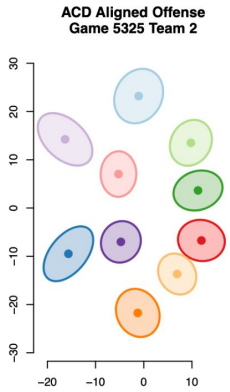

Multicamera Offense
Game 5325 Team 1

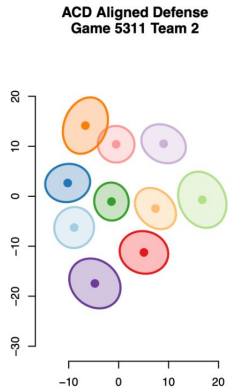

Multicamera Offense
Game 5325 Team 2

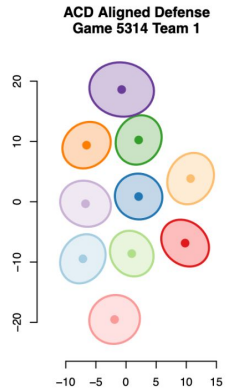

Multicamera Defense
Game 5311 Team 2


Multicamera Defense
Game 5314 Team 1

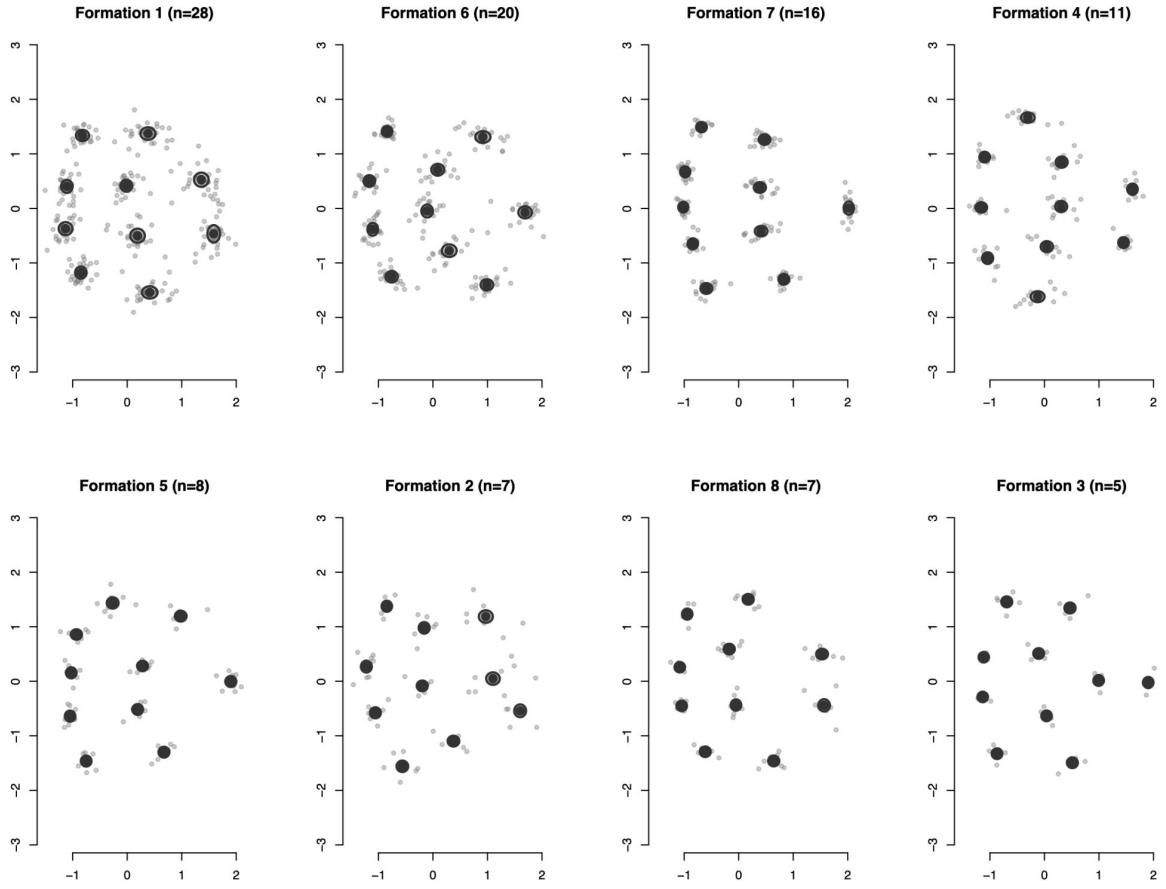

ACD Aligned Offense
Game 5325 Team 1


ACD Aligned Offense
Game 5325 Team 2


ACD Aligned Defense
Game 5311 Team 2


ACD Aligned Defense
Game 5314 Team 1

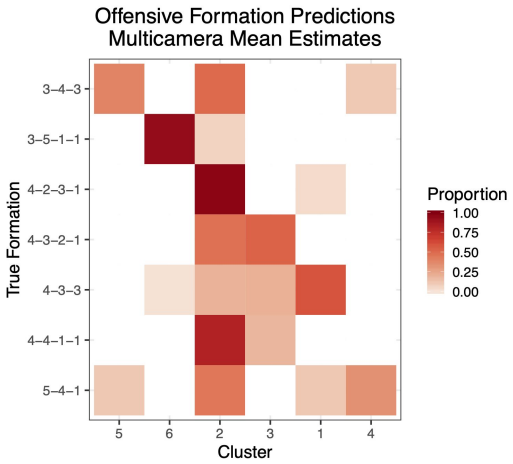
# Results

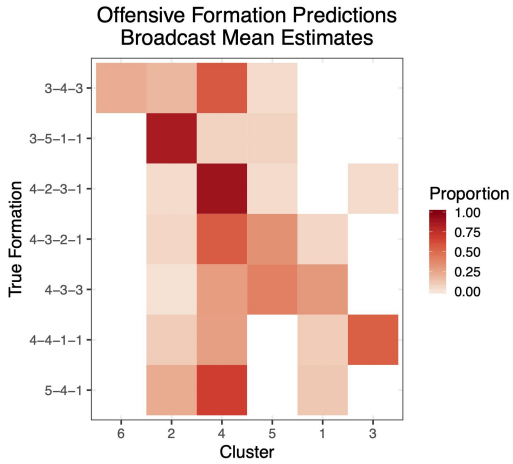

Formation 1 (n=28)
Formation 6 (n=20)
Formation 7 (n=16)
Formation 4 (n=11)
Formation 5 (n=8)
Formation 2 (n=7)
Formation 8 (n=7)
Formation 3 (n=5)

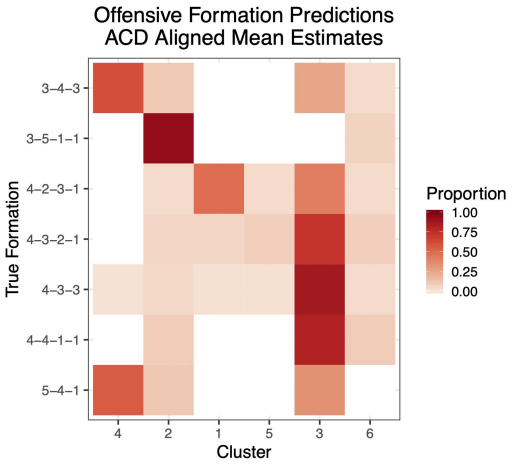
# Results

- Fit mixture models to full multicamera data, broadcast mean calculated using average differences, and ACD aligned data.
- Compare clustered observations with expert-provided formation labels.

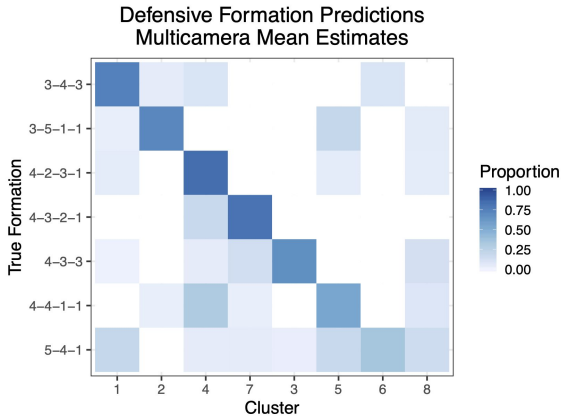
# Results

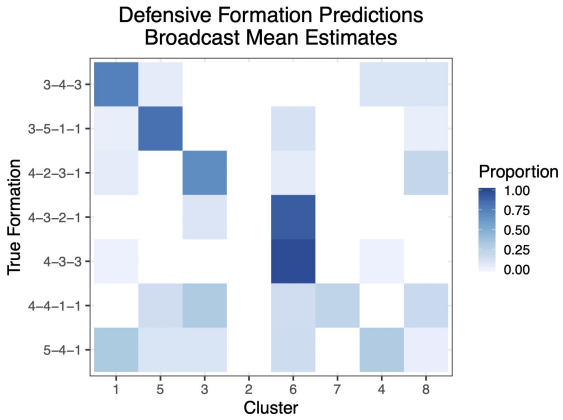

Offensive Formation Predictions
Multicamera Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
5 6 2 3 1 4
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00

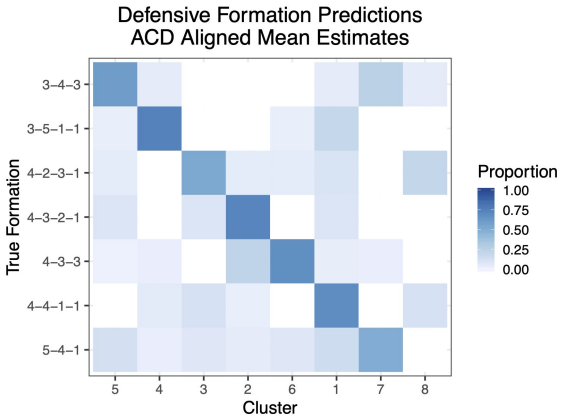

Offensive Formation Predictions
Broadcast Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
6 2 4 5 1 3
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00


Offensive Formation Predictions
ACD Aligned Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
4 2 1 5 3 6
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00

# Results


Defensive Formation Predictions
Multicamera Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
1 2 4 7 3 5 6 8
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00


Defensive Formation Predictions
Broadcast Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
1 5 3 2 6 7 4 8
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00


Defensive Formation Predictions
ACD Aligned Mean Estimates
True Formation
3–4–3
3–5–1–1
4–2–3–1
4–3–2–1
4–3–3
4–4–1–1
5–4–1
5 4 3 2 6 1 7 8
Cluster
Proportion
1.00
0.75
0.50
0.25
0.00

# Summary

- Minimize atomic configuration distance to align frames
- Use MEDP to align possessions and create pseudo observations
- Use Gaussian mixture model to cluster and identify canonical formations

# Acknowledgements

# Acknowledgements


Natural Sciences and Engineering Research Council of Canada

Conseil de recherches en sciences naturelles et en génie du Canada

Canada

SFU SIMON FRASER UNIVERSITY

STATS

SPORT LOGiQ

Thank you!