Spatio-Directional Clustering for Soccer Actions Pattern Discovery

Joaquin Garay

December 15th, 2025

Abstract

We study unsupervised discovery of spatio-directional action structures in professional soccer by clustering each on-ball action using its start location on the pitch and its direction (angle). We propose a Bregman soft-clustering approach for exponential-family mixtures, reformulating EM in expectation-parameter (dual) space so that M-steps reduce to Bregman centroid updates (weighted averages of sufficient statistics). This yields closed-form parameter updates for a bivariate Gaussian location model and a von Mises directional model, avoiding the numerical optimization typically required for von Mises concentration by using accurate analytic inverses of the mean resultant length. We instantiate this framework in two modeling strategies: (I) a Two-layer hierarchical mixture that clusters locations first and then directions conditional on location clusters, and (II) a One-shot mixture that clusters location and direction jointly via a shared latent component. We benchmark multiple EM variants (classical vs. Bregman M-steps, optional C-step hard reclassification, and several initializations) on StatsBomb event data from the 64 matches of the 2018 FIFA World Cup across ten action types. Empirically, the One-shot scheme dominates the Two-layer alternative on BIC/ICL and computational efficiency across most configurations, while Bregman M-steps provide consistent speedups and often modest fit improvements; scaling experiments further indicate linear runtime growth in sample size.

1 Introduction

Modern soccer analytics relies heavily on data-driven representations of on-ball actions. Event and tracking datasets now record where an action starts and how it is played, and these patterns feed into downstream models for possession value, expected goals, and tactical profiling. A central question is therefore how to learn a compact set of "action templates" that capture heterogeneity across the pitch and expose interpretable tactical structure.

Mixture models are a natural tool for this task. Decroos's SoccerMix [1] introduced a bespoke Gaussian–von Mises mixture to represent passes and other actions, and showed that soft cluster memberships provide richer inputs to valuation models than coarse hand-crafted action types. However, SoccerMix chooses a particular hierarchical approach without enough mathematical fundaments and relies on standard EM updates, which for circular data require numerical optimization or an approximation of the von Mises parameters in each M-step. This limits both the theoretical validity of the model class and the and interpretability of the results.

In parallel, work on Bregman clustering has shown that both k-means-style hard clustering and mixture-model EM admit a unifying treatment in terms of Bregman divergences [2]. For exponential-family distributions, Banerjee et al. established a bijection between log-likelihoods and Bregman divergences, and proved that cluster centers in expectation-parameter space are simply weighted averages of sufficient statistics. This implies that the EM M-step for an exponential-family mixture can be expressed as a Bregman centroid computation in the dual coordinates. Conceptually, this is a textbook result; in practice, it has not yet been systematically instantiated and evaluated for spatio-directional action models in soccer.

The present work brings these two strands together. We consider the problem of clustering on-ball actions by their

start location on the pitch and direction, encoded as a polar angle. We model location with a bivariate Gaussian and direction with a von Mises distribution, both in exponential-family form. By working in expectation-parameter space, we recover closed-form M-steps for both components: the Gaussian part reduces to the usual updates for mean and covariance, while the von Mises part uses analytic mappings between expectation and natural parameters together with accurate closed-form approximations for the inverse mean resultant length. This removes a numerical optimizer from the M-step of the directional component and makes the algorithm more robust and predictable.

The remainder of the article is organized as follows. Chapter 2 provides a detailed explanation of mixture models, exponential families, Bregman divergences, and the EM algorithm. Chapter 3 introduces the Gaussian and von Mises distributions and their main properties. Chapter 4 presents the proposed models, the different model variations, and the performance measures. Chapter 5 reports the statistical results.

2 Bregman Clustering Framework

2.1 Mixture Models and EM Algorithm

Consider a mixture distribution

$$p(\mathbf{x}|\gamma) = \sum_{j=1}^{K} \pi_j \, f(\mathbf{x}|\boldsymbol{\theta}_j), \tag{1}$$

where f is a probability density function parametrized by $\boldsymbol{\theta}$ and γ is the set of parameters $\gamma := \{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^{K}$ with $\sum_j \pi_j = 1$. Each distribution component corresponds to a cluster in a soft clustering procedure. It is referred to as "soft" because each data point x_i has a non-zero probability π_j of belonging to each cluster or component C_j.

The weights π_j could be thought of as the probability of a latent random variable Y with categorical distribution, $\pi_j := \text{Prob}(Y = j)$, hence mixture models are an case of discrete latent-variable models.

To fit such models, we want to maximize the data log-likelihood $\log p(\mathbf{x}|\gamma)$ by changing γ. The Expectation-Maximization (EM) algorithm does this job for latent-variable models by maximizing the so-called *evidence lower bound* $\mathcal{L}$. For any distribution q we can write

$$\log p(\mathbf{x}|\gamma) \overset{(a)}{=} \int q(y) \, \log \frac{p(\mathbf{x}, y|\gamma)}{q(y)} \, dy - \int q(y) \, \log \frac{p(y|\mathbf{x}, \gamma)}{q(y)} \, dy, \tag{2}$$

$$= \mathcal{L}(q, \gamma) + D_{KL}[q(y) : p(y|\mathbf{x}, \gamma)] \tag{3}$$

where the first term on the RHS in (a) is the lower bound $\mathcal{L}(q, \gamma)$ and the second term is a Kullback-Leibler divergence. Since any divergence is non-negative, is clear that $\log p(\mathbf{x}|\gamma) \geq \mathcal{L}(q, \gamma)$. In the case of mixture models, we pick $q(y) = p(y|\mathbf{x}, \gamma^{old})$ and replace it into the functional $\mathcal{L}$

$$\mathcal{L}(q_{\gamma^{old}}, \gamma) = \int p(y|\mathbf{x}, \gamma^{old}) \, \log p(\mathbf{x}, y|\gamma) \, dy \, - \int p(y|\mathbf{x}, \gamma^{old}) \log p(y|\mathbf{x}, \gamma^{old}) \, dy \tag{4}$$

$$= \mathbb{E}_Y \left[\log p(\mathbf{x}, y|\gamma) | \mathbf{x}, \gamma^{old} \right] + H \left[p(y|\mathbf{x}, \gamma^{old}) \right]. \tag{5}$$

First term of Equation 5 is the expected complete-data log-likelihood conditioned to the old parametrization, sometimes called $Q(\gamma, \gamma^{old})$ and the second the term is the negative entropy of distribution $p(y|\mathbf{x}, \gamma^{old})$, independent of γ.

EM alternately maximizes $\mathcal{L}$ over one of the inputs $q_{\gamma^{old}}$ or γ while fixing the other:

- **E-Step "Compute the expectation"**: To be able to compute $Q(\gamma, \gamma^{old})$ we need to find the law of the expectation, which is the posterior $p(y|\mathbf{x}, \gamma^{old})$. By Bayes rule, let's define the responsibilities r_{ij} as

$$r_{ij} := p(y_j|\mathbf{x}_i, \gamma) = \frac{p(y_j|\gamma) \, p(\mathbf{x}_i|y_j, \gamma)}{p(\mathbf{x}_i|\gamma)} = \frac{\pi_j \, f(\mathbf{x}_i|\boldsymbol{\theta}_j)}{\sum_{h=1}^{K} \pi_h f(\mathbf{x}_i|\boldsymbol{\theta}_h)}, \tag{6}$$

  so that

$$Q(\gamma, \gamma^{old}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \log \left(\pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j) \right) \cdot r_{ij}^{old} \tag{7}$$

- **M-step "Maximize the expectation"**:

$$\gamma^{old} \leftarrow \arg\max_{\gamma} Q(\gamma, \gamma^{old}) \tag{8}$$

Then, use the vector γ as the old vector γ_0 in a new iteration of the algorithm until convergence.

Additionally, when considering the full dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, the M-step can be reformulated as $2K$ distinct maximization problems. Specifically, the maximization of $Q(\gamma, \gamma^{old})$ can be expressed as

$$Q(\gamma, \gamma^{old}) = \sum_{i=1}^{N}\sum_{j=1}^{K} \log f(\mathbf{x}_i|\boldsymbol{\theta}_j) \cdot r_{ij}^{old} + \sum_{i=1}^{N}\sum_{j=1}^{K} \log \pi_j \cdot r_{ij}^{old}. \tag{9}$$

For all $j \in \{1, \ldots, K\}$, this leads to two subproblems:

$$\pi_j \leftarrow \arg\max_{\pi} \sum_{i=1}^{N} \log \pi \cdot r_{ij}^{old} \quad \text{and} \quad \boldsymbol{\theta}_j \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(\mathbf{x}_i|\boldsymbol{\theta}) \cdot r_{ij}^{old} \tag{10}$$

The update for π_j admits a closed-form solution $\pi_j = \frac{1}{N}\sum_i r_{ij}$, whereas the maximization with respect to $\boldsymbol{\theta}_j$ generally lacks a closed-form solution and thus requires numerical methods. However, this challenge can be addressed using a geometrical approach, which will be introduced in the following section.

The EM algorithm is guaranteed to converge to a local maximum (See [3] for detailed reference.) The algorithm can be outlined as in Algorithm 1.

Algorithm 1 Expectation-Maximization for mixture models

Require: Input data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$.
Initialize $\gamma = \{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^K$.
while not convergence **do**
  E-Step: Compute the posteriors r_{ij}.
  for $i = 1$ **to** N **do**
    for $j = 1$ **to** K **do**
      $r_{ij} \leftarrow \pi_j \cdot f(\mathbf{x}_i|\boldsymbol{\theta}_j) / \sum_{h=1}^{K} \pi_h \, f(\mathbf{x}_i|\boldsymbol{\theta}_h)$
    end for
  end for
  M-Step: Maximize $\mathbb{E}_Y\left[\log p(\mathbf{x}, y|\gamma)|\mathbf{x}, \gamma^{old}\right]$.
  for $j = 1$ **to** K **do**
    $\pi_j \leftarrow \frac{1}{N}\sum_{i=1}^{N} r_{ij}$
    $\boldsymbol{\theta}_j \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(\mathbf{x}_i|\boldsymbol{\theta}) \, r_{ij}$
  end for
end while
return Maximum likelihood estimate parameters $\gamma = \{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^K$

2.2 Exponential Family and Bregman Divergence

The exponential family is a class of parametric probability distributions characterized by a specific functional form. In particular, their probability density function can be written as

$$f(\mathbf{x}|\boldsymbol{\theta}) = \exp\{\langle \mathbf{t}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}) + k(\mathbf{x})\} \tag{11}$$

where $\mathbf{t} := \mathbf{t}(\mathbf{x})$ is the sufficient statistic (vector), $\boldsymbol{\theta}$ is the natural parameter of the distribution, $\psi(\boldsymbol{\theta})$ is the log-partition function (also known as the cumulant or free energy function), and $k(\mathbf{x})$ is a base measure function.

Since $\psi(\boldsymbol{\theta})$ is strictly convex [2], we can compute its Legendre conjugate function ϕ as

$$\boldsymbol{\eta} := \nabla\psi(\boldsymbol{\theta}) \quad \text{and} \quad \phi(\boldsymbol{\eta}) := \sup_{\boldsymbol{\theta}}\{\langle\boldsymbol{\eta}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta})\} \tag{12}$$

On the other hand, a Bregman divergence is defined as

$$D_\phi[\mathbf{v} : \mathbf{v}_0] := \phi(\mathbf{v}) - \phi(\mathbf{v}_0) - \langle\nabla\phi(\mathbf{v}_0), \mathbf{v} - \mathbf{v}_0\rangle \tag{13}$$

for any strictly convex and differentiable function ϕ and some vectors $\mathbf{v}$ and $\mathbf{v}_0$.

Proposition 2.1. *Consider the log-likelihood of an exponential family distribution with natural parameter $\boldsymbol{\theta}$. It can be rewritten as:*

$$\log f(\mathbf{x}|\boldsymbol{\theta}) = -D_\phi[\mathbf{t} : \boldsymbol{\eta}] + \phi(\mathbf{t}) \tag{14}$$

Proof.

$$\begin{aligned}
\langle\boldsymbol{\theta}, \mathbf{t}\rangle - \psi(\boldsymbol{\theta}) &= \langle\boldsymbol{\theta}, \mathbf{t}\rangle - \psi(\boldsymbol{\theta}) + \langle\boldsymbol{\eta}, \boldsymbol{\theta}\rangle - \langle\boldsymbol{\eta}, \boldsymbol{\theta}\rangle \\
&= \langle\boldsymbol{\eta}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) + \langle\mathbf{t} - \boldsymbol{\eta},\, \boldsymbol{\theta}\rangle \\
&= \phi(\boldsymbol{\eta}) + \langle\mathbf{t} - \boldsymbol{\eta}, \nabla\phi(\boldsymbol{\eta})\rangle + \phi(\mathbf{t}) - \phi(\mathbf{t}) \\
&= -D_\phi[\mathbf{t} : \boldsymbol{\eta}] + \phi(\mathbf{t})
\end{aligned}$$

□

This demonstrates that maximizing the log-likelihood of an exponential family distribution is equivalent to minimizing the Bregman divergence induced by the Legendre-conjugate function D_ϕ.

2.3 Cluster Center Theorem and Revisited EM Algorithm

Let $C = \{\mathbf{v}_i\}_{i=1}^N$ be a sample of N data points. We seek a representative point $\boldsymbol{\eta}$ that is as "close" as possible to all members of C, weighted by a sample weight vector $\mathbf{w}$. To quantify this closeness, we minimize the average Bregman divergence from each point in the cluster to $\boldsymbol{\eta}$, defined as:

$$D_\phi[C : \boldsymbol{\eta}] = \sum_{\mathbf{x}_i \in C} w_i\, D_\phi[\mathbf{v}_i : \boldsymbol{\eta}]. \tag{15}$$

The point $\boldsymbol{\eta}$ that minimizes this expression is called the ϕ-center of the cluster C, with respect to divergence D_ϕ. [4]

Theorem 2.2. *(Cluster center) The ϕ-center of cluster C is given by*

$$\boldsymbol{\eta}_C = \frac{\sum_{\mathbf{v}_i \in C} w_i\, \mathbf{v}_i}{\sum_{\mathbf{v}_i \in C} w_i} \tag{16}$$

for any strictly convex and differentiable function ϕ.

Proof. We begin with the definition of the Bregman divergence:

$$\min_{\boldsymbol{\eta}} D_\phi[C:\boldsymbol{\eta}] = \min_{\boldsymbol{\eta}} \sum_{\mathbf{v}_i \in C} w_i \left[\phi(\mathbf{v}_i) - \phi(\boldsymbol{\eta}) - \langle \nabla\phi(\boldsymbol{\eta}), \mathbf{v}_i - \boldsymbol{\eta}\rangle\right].$$

Then, first-order optimality condition yields

$$\frac{\partial}{\partial \boldsymbol{\eta}} D_\phi[C:\boldsymbol{\eta}] = \sum_{\mathbf{v}_i \in C} w_i \left[-\nabla\phi(\boldsymbol{\eta}) - \nabla\nabla\phi(\boldsymbol{\eta}) \cdot (\mathbf{v}_i - \boldsymbol{\eta}) + \nabla\phi(\boldsymbol{\eta})\right] = 0$$

$$\sum_{\mathbf{v}_i \in C} -\nabla\nabla\phi(\boldsymbol{\eta}) \cdot w_i(\mathbf{v}_i - \boldsymbol{\eta}) = 0$$

Since $\nabla\nabla\phi(\boldsymbol{\eta})$ is positive definite,

$$\sum_{\mathbf{v}_i \in C} w_i\, \mathbf{v}_i - \boldsymbol{\eta} \sum_{\mathbf{v}_i \in C} w_i = 0 \quad \Rightarrow \quad \boldsymbol{\eta} = \frac{\sum_{\mathbf{v}_i \in C} w_i\, \mathbf{v}_i}{\sum_{\mathbf{v}_i \in C} w_i}$$

□

Remark 2.3. The minimization is taken with respect to the second argument of D_ϕ, and it is important to note that Bregman divergences are not necessarily convex in that argument [2].

Proposition 2.4. *Consider a mixture model of exponential family distributions, with density given by*

$$p(\mathbf{x}|\gamma) = \sum_{j=1}^{K} \pi_j \exp\{\langle \mathbf{t}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) + k(\mathbf{x})\}. \tag{17}$$

The M-step for updating the parameters $\boldsymbol{\theta}$ in Algorithm 1, formerly expressed as

$$\boldsymbol{\theta}_j \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(\mathbf{x}_i|\boldsymbol{\theta}) \cdot r_{ij}, \quad \forall j \in \{1, \ldots, k\} \tag{18}$$

can be reduced to the following closed-form expression

$$\boldsymbol{\eta}_j \leftarrow \frac{\sum_{i=1}^{N} r_{ij} \cdot \mathbf{t}_i}{\sum_{i=1}^{N} r_{ij}}. \tag{19}$$

where $\boldsymbol{\eta}_j$ is the Legendre dual coordinates of $\boldsymbol{\theta}_j$, also known as the expectation parameter $\mathbb{E}[\mathbf{t}]$.

Proof.

$$\begin{aligned}
\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(\mathbf{x}_i|\boldsymbol{\theta}) \cdot r_{ij} &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \{\langle \boldsymbol{\theta}, \mathbf{t}_i\rangle - \psi(\boldsymbol{\theta}) + k(\mathbf{x}_i)\} \cdot r_{ij} \\
&= \max_{\boldsymbol{\eta}} \sum_{i=1}^{N} \{-D_\phi[\mathbf{t}_i:\boldsymbol{\eta}] + \phi(\mathbf{t}_i)\} \cdot r_{ij} \\
&= \min_{\boldsymbol{\eta}} \sum_{i=1}^{N} D_\phi[\mathbf{t}_i:\boldsymbol{\eta}] \cdot r_{ij} - \text{constant}
\end{aligned}$$

By Theorem 2.2 it follows that

$$\boldsymbol{\eta}_j = \frac{\sum_{i=1}^{N} r_{ij}\, \mathbf{t}_i}{\sum_{i=1}^{N} r_{ij}}$$

□

This derivation provides a straightforward method for computing the parameter updates. However, it requires working in the dual coordinate system. If the conjugate function ϕ cannot be computed in closed form and must instead be derived via the Legendre transformation $\phi(\boldsymbol{\eta}) = \sup_{\boldsymbol{\theta}}\{\langle \boldsymbol{\eta}, \boldsymbol{\theta} \rangle - \phi(\boldsymbol{\theta})\}$, then we are essentially faced with the same computational difficulty as in the original EM algorithm, and no practical advantage is gained. [2]

3 Inference of actions

We model two inputs: (I) the action's start location on the pitch and (II) its direction, measured as a polar angle relative to the positive x-axis. We define distinct distributions for the two input types: the first is modelled by a bivariate Gaussian and the latter as a von Mises.

3.1 Location Modeling

Proposition 3.1. *A d-dimensional multivariate Gaussian distribution*

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\,\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{20}$$

can be expressed in the exponential-family form with

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ -\frac{1}{2}\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\}\end{pmatrix}, \qquad \mathbf{t}(\mathbf{x}) = \begin{pmatrix}\mathbf{x}\\ \mathrm{vec}\{\mathbf{x}\mathbf{x}^\top\}\end{pmatrix}, \qquad \psi(\boldsymbol{\theta}) = \frac{1}{2}\left[d\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right], \tag{21}$$

where $\mathrm{vec}\{\cdot\}$ *is a vectorized operation that stacks every row in a column vector.*

Proof. Rearrange f as

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp\left\{\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(d\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}. \tag{22}$$

and note that the second term can be written as $-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathrm{tr}\left[-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^\top\right] = \langle -\frac{1}{2}\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\}, \mathrm{vec}\{\mathbf{x}\mathbf{x}^\top\}\rangle$. □

Corollary 3.2. *Expectation parameter* $\boldsymbol{\eta}$ *and the dual log-partition conjugate function* $\phi(\boldsymbol{\eta})$ *are read as*

$$\boldsymbol{\eta} = \mathbb{E}\left[\mathbf{t}\right] = \begin{pmatrix}\boldsymbol{\mu}\\ \mathrm{vec}\{\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\}\end{pmatrix}, \qquad \phi(\boldsymbol{\eta}) = -\frac{1}{2}\left(d + d\ln(2\pi) + \ln|\boldsymbol{\Sigma}|\right) \tag{23}$$

Proof. It is easy to see that $\mathbb{E}\left[x_i x_j\right] = \Sigma_{ij} + \mu_i\mu_j$, for all $(i,j) \in \{1,\dots,d\}\times\{1,\dots,d\}$.

$$\begin{aligned}
\phi(\boldsymbol{\eta}) &= \langle\boldsymbol{\eta},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta})\\
&= \left\langle \begin{pmatrix}\boldsymbol{\mu}\\ \mathrm{vec}\{\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ -\frac{1}{2}\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}\}\end{pmatrix}\right\rangle - \frac{1}{2}\left(d\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\\
&= \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\mathrm{tr}\left[(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{\Sigma}^{-1}\right] - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\left(d\ln(2\pi) + \ln|\boldsymbol{\Sigma}|\right)\\
&= -\frac{1}{2}\mathrm{tr}[\mathbf{I}_d] - \frac{1}{2}\left(d\ln(2\pi) + \ln|\boldsymbol{\Sigma}|\right)
\end{aligned}$$

□

With these parameters defined, it is of our interest to have the dual divergence of log-partition function, which is equivalent to the Kullback-Leibler divergence of the distribution

$$D_\phi[\boldsymbol{\eta} : \boldsymbol{\eta}_0] = D_{KL}[p : p_0] = \frac{1}{2}\left(\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}|} + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \mathrm{tr}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}] - d\right) \tag{24}$$

Proof.

$$
\begin{aligned}
D_\phi[\boldsymbol{\eta} : \boldsymbol{\eta}_0] &= \phi(\boldsymbol{\eta}) - \phi(\boldsymbol{\eta}_0) - \langle \boldsymbol{\theta}_0, \boldsymbol{\eta} - \boldsymbol{\eta}_0 \rangle \\
&= -\frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\ln|\boldsymbol{\Sigma}_0| - \left\langle \left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0,\, -\frac{1}{2}\mathrm{vec}\{\boldsymbol{\Sigma}_0^{-1}\} \right), \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0,\, \mathrm{vec}\left\{\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top\right\}\right) \right\rangle \\
&= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}|} - \boldsymbol{\mu}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0 + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\right] \\
&= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}|} - \boldsymbol{\mu}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\right] + \frac{1}{2}\boldsymbol{\mu}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}_0\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}d \\
&= \frac{1}{2}\left(\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}|} + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \mathrm{tr}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}] - d\right)
\end{aligned}
$$

□

3.2 Direction Modeling

The von Mises is a probability distribution defined on a unit circle. It has a probability density function

$$f(x|\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(x-\mu)\}. \tag{25}$$

with $I_0(\kappa)$ the Modified Bessel function of first kind.

Proposition 3.3. *Let $X \in (-\pi, \pi]$ be an angular random variable. Its von Mises density with location $\mu \in (-\pi, \pi]$ and concentration $\kappa \geq 0$, Equation 25, belongs to the minimal two-parameter exponential family with*

$$\boldsymbol{\theta} = (\kappa\cos\mu,\ \kappa\sin\mu)^\top, \qquad \mathbf{t}(x) = (\cos x,\ \sin x)^\top, \qquad \psi(\boldsymbol{\theta}) = \ln 2\pi I_0(\|\boldsymbol{\theta}\|) \tag{26}$$

and base measure $k(x) = 0$, where $\|\boldsymbol{\theta}\| = \sqrt{\theta_1^2 + \theta_2^2} = \kappa$.

Proof. Using $\cos(x-\mu) = \cos x\cos\mu + \sin x\sin\mu$,

$$\langle \boldsymbol{\theta}, \mathbf{t}\rangle = \theta_1\cos x + \theta_2\sin x = \kappa(\cos\mu\cos x + \sin\mu\sin x) = \kappa\cos(x-\mu).$$

Therefore

$$\exp\big\{\langle\boldsymbol{\theta},\mathbf{t}\rangle - \psi(\boldsymbol{\theta})\big\} = \exp\{\kappa\cos(x-\mu)\}\big(2\pi I_0(\kappa)\big)^{-1} = f_X(x \mid \mu, \kappa).$$

Finally, $\|\boldsymbol{\theta}\| = \sqrt{(\kappa\cos\mu)^2 + (\kappa\sin\mu)^2} = \kappa$, completing the proof. □

Remark 3.4. $I_v(x)$ is the Modified Bessel function of first kind.

$$I_v(x) = \sum_{k=0}^{\infty}\frac{1}{k!\,\Gamma(k+v+1)}\left(\frac{x}{2}\right)^{2k+v}$$

Remark 3.5. When $\kappa = 0$ the distribution reduces to the uniform density on the circle, and μ is not identifiable.

Corollary 3.6. *Expectation parameter $\boldsymbol{\eta}$ and the dual log-partition conjugate function $\phi(\boldsymbol{\eta})$ are describe as*

$$\eta(\mu,\kappa) = (A(\kappa)\cos\mu, A(\kappa)\sin\mu)^\top, \qquad \phi(\mu,\kappa) = \kappa A(\kappa) - \ln 2\pi I_0(\kappa) \tag{27}$$

with $A(\kappa) = I_1(\kappa)/I_0(\kappa)$.

Proof. Using the fact that $\boldsymbol{\eta} = \mathbb{E}[\mathbf{t}]$, we compute the first component η_1.

$$\begin{aligned}
\eta_1 &= \mathbb{E}[\cos x] \\
&= \frac{1}{2\pi I_0(\kappa)} \int_{-\pi}^{\pi} \cos x \cdot \exp\{\kappa \cos(x-\mu)\}\, dx \\
&= \frac{1}{2\pi I_0(\kappa)} \int_{-\pi}^{\pi} \cos(x+\mu) \cdot \exp\{\kappa \cos x\}\, dx \\
&= \frac{\cos\mu}{2\pi I_0(\kappa)} \int_{-\pi}^{\pi} \cos x \cdot \exp\{\kappa \cos x\}\, dx - \frac{\sin\mu}{2\pi I_0(\kappa)} \int_{-\pi}^{\pi} \sin x \cdot \exp\{\kappa \cos x\}\, dx \\
&\overset{(a)}{=} \frac{\cos\mu}{2\pi I_0(\kappa)} \int_{-\pi}^{\pi} \cos x \cdot \exp\{\kappa \cos x\}\, dx \\
&\overset{(b)}{=} \frac{\cos\mu}{2\pi I_0(\kappa)} \cdot 2\pi I_1(\kappa) \\
&= A(\kappa)\cos\mu.
\end{aligned}$$

In (a) we use the property of odd function integration and in (b) we use even function integration property along a known equality of the modified Bessel function: $I_v(\kappa) = \frac{1}{\pi}\int_0^{\pi} \cos(vx)\exp\{\kappa\cos x\}\, dx$. The result of the second component η_2 follows the same argument.
The convex function ϕ can be computed as

$$\begin{aligned}
\phi(\mu,\kappa) &= \langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle - \psi(\boldsymbol{\theta}) \\
&= \kappa\cos\mu \cdot A(\kappa)\cos\mu + \kappa\sin\mu \cdot A(\kappa)\sin\mu - \ln 2\pi I_0(\kappa) \\
&= \kappa A(\kappa) - \ln 2\pi I_0(\kappa)
\end{aligned}$$

□

Remark 3.7. We can write the function in function of $\boldsymbol{\eta}$ as

$$\phi(\boldsymbol{\eta}) = A^{-1}(\|\boldsymbol{\eta}\|) \cdot \|\boldsymbol{\eta}\| - \ln 2\pi I_0(A^{-1}(\|\boldsymbol{\eta}\|))$$

Corollary 3.8. *Dual divergence of von Mises distribution*

$$D_\phi[\boldsymbol{\eta} : \boldsymbol{\eta}_0] = KL[p : p_0] = \kappa A(\kappa) + \ln \frac{I_0(\kappa_0)}{I_0(\kappa)} - \kappa_0 A(\kappa)\cos(\mu_0 - \mu) \tag{28}$$

Proof.

$$\begin{aligned}
D_\phi[\boldsymbol{\eta} : \boldsymbol{\eta}_0] &= \phi(\boldsymbol{\eta}) - \phi(\boldsymbol{\eta}_0) - \langle \boldsymbol{\theta}_0, \boldsymbol{\eta} - \boldsymbol{\eta}_0 \rangle \\
&= \kappa A(\kappa) - \ln 2\pi I_0(\kappa) - \kappa_0 A(\kappa_0) + \ln 2\pi I_0(\kappa_0) \\
&\quad - \langle (\kappa_0\cos\mu_0,\ \kappa_0\sin\mu_0), (A(\kappa)\cos\mu - A(\kappa_0)\cos\mu_0,\ A(\kappa)\sin\mu - A(\kappa_0)\sin\mu_0) \rangle \\
&= \kappa A(\kappa) - \kappa_0 A(\kappa_0) + \ln \frac{I_0(\kappa_0)}{I_0(\kappa)} + \kappa_0 A(\kappa_0)\,(\cos^2\mu_0 + \sin^2\mu_0) - \kappa_0 A(\kappa)\,(\cos\mu_0\cos\mu + \sin\mu_0\sin\mu) \\
&= \kappa A(\kappa) + \ln \frac{I_0(\kappa_0)}{I_0(\kappa)} - \kappa_0 A(\kappa)\cos(\mu_0 - \mu)
\end{aligned}$$

□

There remains one computational step when working in dual coordinates: inverting the mean resultant length,

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}, \qquad A : (0, \infty) \to (0, 1). \tag{29}$$

Since A is strictly increasing, the inverse A^{-1} exists, and we can recover κ from $R = \|\boldsymbol{\eta}\|$ with a 1-D root-finder. For speed, we can use closed-form approximations. A simple yet efficient candidate that was developed by Banerjee et al. [5] and use by SoccerMix is read as

$$A^{-1}(R) \approx \frac{R\,(2 - R^2)}{1 - R^2}, \tag{30}$$

although a more accurate piecewise approximation was designed years before by Best and Fisher in [6]

$$A^{-1}(R) \approx \begin{cases} \frac{5}{6}R^5 + R^3 + 2R, & R < 0.53, \\ 1.39R - 0.4 + \dfrac{0.43}{1 - R}, & 0.53 \le R < 0.85, \\ \dfrac{1}{R^3 - 4R^2 + 3R}, & R \ge 0.85. \end{cases} \tag{31}$$

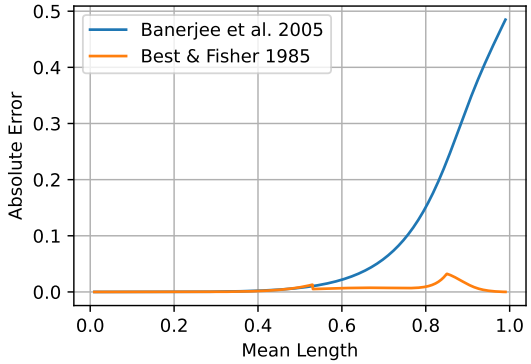


Figure 1: Approximation absolute errors on the inverse of the Mean Resultant Length function

4 Experimental Setup

Two different models are proposed for the task of modeling location and direction of on-ball actions, namely the Two-layer Scheme and One-shot Scheme. Each one allows the different M-steps of the EM algorithm discussed before, and also the optional C-step, another EM modification for clustering purpose that will be discussed later.

An extra layer of combinations is be consider when we explore the different initialization methods for the models.

4.1 Two-layer Scheme

Each observation $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ records the start location and direction of a ball action. The location data $(x_{i1}, x_{i2}) \in [0, 105] \times [0, 68]$ (meters) gives field coordinates, and the direction data (x_{i3}, x_{i4}) represents the cosine and sine of the angle measured from the positive x-axis, respectively.

We use a two-layer clustering scheme.

- Layer 1 clusters locations with a 2-D Gaussian mixture fitted by EM, yielding the mixture parameter set γ^{loc} and the location-cluster responsibilities r_{ij}^{loc}.
- Layer 2 then fits, for each location cluster j, a von Mises mixture to the direction variable, running EM with sample weights equal to r_{ij}^{loc}. Thus each observation contributes to cluster j in proportion to its location-level posterior. This hierarchy separates spatial structure from directional tendencies while sharing information through the responsibilities.
- The number of directional clusters is flexible to depend to its assigned location cluster.

Proposition 4.1. *The probability density function of such a model will read*

$$p_{\text{two-layer}}(\mathbf{x}_i) = \sum_{j=1}^{K^{loc}} \pi_j \, f_{loc}\left(\mathbf{x}_{i,loc} | \boldsymbol{\theta}_j\right) \cdot \sum_{k=1}^{K_j^{dir}} \omega_{j,k} \, f_{dir}\left(\mathbf{x}_{i,dir} | \boldsymbol{\theta}_{j,k}\right), \tag{32}$$

Proof. Consider location and direction random variables $\mathbf{x}_{loc}$ and $\mathbf{x}_{dir}$, respectively. Furthermore, let y be the latent random variable of location and z the latent random variable of direction (dependent on location).

$$\begin{aligned}
p(\mathbf{x}_{loc}, \mathbf{x}_{dir}) &= \sum_y p(\mathbf{x}_{loc}, \mathbf{x}_{dir} | y)\, p(y), && (33)\\
&= \sum_y p(\mathbf{x}_{dir} | \mathbf{x}_{loc}, y) \;\cdot p(\mathbf{x}_{loc} | y)\, p(y), && (34)\\
&= \sum_y \left(\sum_{z|y} p(\mathbf{x}_{dir} | \mathbf{x}_{loc}, y, z)\, p(z|y) \right) \; p(\mathbf{x}_{loc} | y)\, p(y), && (35)\\
&= \sum_y \left(\sum_{z|y} f_{dir}(\mathbf{x}_{dir} | \boldsymbol{\theta}_{y,z})\, \omega_{y,z} \right) \; f_{loc}(\mathbf{x}_{loc} | \boldsymbol{\theta}_y)\, \pi_y, && (36)
\end{aligned}$$

where we assumed $(z \perp \mathbf{x}_1 \,|\, y)$ in (35) and $(\mathbf{x}_2 \perp \mathbf{x}_1 \,|\, y, z)$ in (36). Replacing the latent variables y and $z|y$ to the categorical support $y \in \{1, \ldots, K^{loc}\}$ and $z \in \{1, \ldots, K^{dir}(y)\}$

$$p(\mathbf{x}_{loc}, \mathbf{x}_{dir}) = \sum_{j=1}^{K^{loc}} \pi_j \, f_{loc}\left(\mathbf{x}_{i,loc} | \boldsymbol{\theta}_j\right) \cdot \sum_{k=1}^{K_j^{dir}} \omega_{j,k} \, f_{dir}\left(\mathbf{x}_{i,dir} | \boldsymbol{\theta}_{j,k}\right), \tag{37}$$

□

Remark 4.2. Assuming conditional independency between $\mathbf{x}_{loc}$ and $\mathbf{x}_{dir}$ doesn't mean they are marginally independent; when the latent parents y, z are unknown, $\mathbf{x}_{loc}$ gives information about y, which gives information about $\mathbf{x}_{dir}$. Conditional independence means that locally the two features look independent.

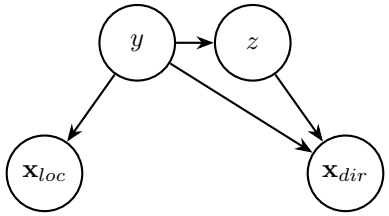


Figure 2: Two-layer model Bayesian network

4.2 Convex Combination of Bregman Divergences One-shot Scheme

The following model was inspired by [2], which aims to eliminate the hierarchical step of modeling the two different features. We can make use of Bregman divergence and exponential family properties again to implement simultaneous multidimensional cluster approach.

Theorem 4.3. ***(Barnejee et. al.)*** *There is a bijection between regular Exponential Family distributions and regular Bregman divergences.*

Proposition 4.4. *Consider ϕ_{Gauss} and ϕ_{vM} the Legendre-conjugate of the log-partition function of a Gaussian and von Mises distribution, respectively. Then, a convex combination of both functions induces an instance of an Exponential Family distribution expressed as*

$$\phi^*(\mathbf{t}) = \alpha\,\phi_{Gauss}(\mathbf{t}_{loc}) + \beta\,\phi_{vM}(\mathbf{t}_{dir}) \Rightarrow p^*(\mathbf{x}) = [f_{Gauss}(\mathbf{x}_{loc}|\boldsymbol{\theta}_{Gauss})]^{\alpha}\,[f_{vM}(\mathbf{x}_{dir}|\boldsymbol{\theta}_{vM})]^{\beta} \cdot \mathbf{b}_{\phi^*}(\mathbf{x}) \tag{38}$$

with $\mathbf{b}$ *a normalizing factor.*

Proof. Given that $\mathbf{t}$ is a stacked vector of the sufficient statistics of both Gaussian and von Mises distributions,

$$\begin{aligned} D_{\phi^*}[\mathbf{t} : \mathbf{t_0}] &= \alpha\,\phi_{Gauss}(\mathbf{t}_{loc}) + \beta\,\phi_{vM}(\mathbf{t}_{dir}) - \alpha\,\phi_{gauss}(\mathbf{t}_{0,loc}) - \beta\,\phi_{vM}(\mathbf{t}_{0,dir}) \\ &\quad - \alpha\,\langle \nabla\phi_{gauss}(\mathbf{t}_{0,loc}),\, \mathbf{t}_{loc} - \mathbf{t}_{0,loc}\rangle - \beta\,\langle \nabla\phi_{vM}(\mathbf{t}_{0,dir}),\, \mathbf{t}_{dir} - \mathbf{t}_{0,dir}\rangle \\ &= \alpha\, D_{\phi_{gauss}}[\mathbf{t}_{loc} : \mathbf{t}_{0,loc}] + \beta\, D_{\phi_{vM}}[\mathbf{t}_{dir} : \mathbf{t}_{0,dir}] \end{aligned}$$

Hence

$$\begin{aligned} \exp\{-D_{\phi^*}[\mathbf{t} : \boldsymbol{\eta}]\}\mathbf{b}_{\phi^*}(\mathbf{x}) &= \exp\{-\alpha\, D_{\phi_{Gauss}}[\mathbf{t}_{loc} : \boldsymbol{\eta}_{loc}]\} \cdot \exp\{-\beta\, D_{\phi_{vM}}[\mathbf{t}_{dir} : \boldsymbol{\eta}_{dir}]\}\mathbf{b}_{\phi^*}(\mathbf{x}) \\ &= [f_{Gauss}(\mathbf{x}|\boldsymbol{\theta}_{Gauss})]^{\alpha}\,[f_{vM}(\mathbf{x}|\boldsymbol{\theta}_{vM})]^{\beta} \cdot \mathbf{b}_{\phi^*}(\mathbf{x}) \end{aligned}$$

□

Hence, this convex combination is not capable of capturing the dependencies between both features. However, a mixture of of this distribution $p^*(\mathbf{x})$ will link $\mathbf{x}_{loc}$ and $\mathbf{x}_{dir}$ through a latent variable y, so that each mixture component $p^*(\mathbf{x}_{loc}, \mathbf{x}_{dir}|y) = p(\mathbf{x}_{loc}|y) \cdot p(\mathbf{x}_{dir}|y)$.

Because we want to use an scale-modulated distribution for each feature, these weights α and β enter in conflict with the scale parameter of the distribution, creating an effect of an extra degree of freedom. To solve this problem, we propose to use $\alpha = \beta = 1$, and keep the full parametrization of the distributions.

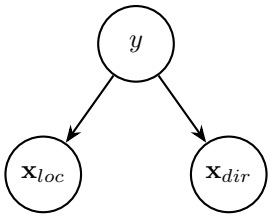


Figure 3: One-shot model Bayesian network

Proposition 4.5. *The probability density function of the "one-shot" model will read*

$$p_{\text{one-shot}}(\mathbf{x}_i) = \sum_{j=1}^{K} \pi_j \, f_{loc}(\mathbf{x}_{i,loc}|\boldsymbol{\theta}_j) \cdot f_{dir}(\mathbf{x}_{i,dir}|\boldsymbol{\theta}_j) \tag{39}$$

Remark 4.6. Similar to the previous model, we assumed $(\mathbf{x}_{loc} \perp \mathbf{x}_{dir} \,|\, y)$

4.3 EM Variations

We compare different EM implementations, making modifications to Algorithm 1.

E-step: The E-step (computation of responsibilities r_{ij}) is identical in all variants.

C-step: The Classification EM (CEM) algorithm is a variant that adds a new C-step to Algorithm 1 such that it maximizes the complete-data log-likelihood instead [7]. It consist in a one-hot encoding of the responsibility matrix $\mathbf{R}$ after the E-step and before the M-step, such that each observation belong to one and only one cluster j.

$$\mathbf{R}^* = \left\{ [r^*_{ij}]_{1\leq i\leq N,\, 1\leq j\leq K} : r^*_{ij} = 1 \text{ if } \arg\max_l r_{il} = j \text{ else } 0,\ \forall i \right\} \tag{40}$$

This allowed a better separation between clusters, closer like a K-means, while maintaining the probabilistic aspect of the model. We evaluate models with and without the C-step modification.

M-step: The *Bregman* variation is to use the dual/expectation coordinates $\boldsymbol{\eta}$ and the results exposed in Proposition 2.4, mapping back to the ordinary parameters for reporting.

The Gaussian distribution has a well-known closed-form M-step in ordinary parameters which is used in the classical EM approach. Both routes are algebraically equivalent.

For the von Mises distribution, the classical EM maximizes the data likelihood in (μ_j, κ_j) with a numerical solver, while the Bregman M-step uses dual coordinates to then map back via $\mu_j = \text{atan2}(\eta_{j,2}, \eta_{j,1})$ and $\kappa_j \approx A^{-1}(\|\boldsymbol{\eta}_j\|)$ of Equation 30 or Equation 31. This removes a numerical optimizer from the M-step, aside from the inexpensive inversion A^{-1}.

4.4 Performance Metrics

Our goal is to find the mixture model that shows the greatest evidence for clustering the dataset $\mathcal{X}$. Biernacki et. al. [8] explained Bayesian Information Criterion is a rough approximation of the integrated log-likelihood over the parameter space of model m with K components. Thus, finding the model m and component number K that maximize this integrated log-likelihood is equivalent, at a practical level, to $(\hat{m}, \hat{K}) = \arg\max_{m,K} BIC(m, K)$,

$$BIC(m, K) := \sum_i \log p(\mathbf{x}_i) - \frac{\nu_{m,K}}{2} \log n, \tag{41}$$

with $\nu_{m,K}$ the number of free parameters in the model. However, since this metric is not reflective on how well the model exhibit the clustering structure of the data, we also include the Integrated Completed Likelihood (ICL) proposed in [8] as an extra metric, which incorporates the complete-data log-likelihood instead

$$ICL(m, K) = \sum_i \log p(\mathbf{x}_i, \hat{y}_i) - \frac{\nu_{m,K}}{2} \log n. \tag{42}$$

Because y_i is latent, we replace it with $\hat{y}_i = \arg\max_j r_{ij}$, a maximum a posteriori estimation. To follow the standard convention, we will aim and report the minimization of $-2 \cdot BIC$ and $-2 \cdot ICL$.

For the One-shot model, the complete-data log-likelihood goes

$$\log p_{\text{one-shot}}(\mathbf{x}_i, y_i) = \sum_{j=1}^{K} y_{ij} \log \{\pi_j \, f_{loc}(\mathbf{x}_{i,loc}|\boldsymbol{\theta}_j) \cdot f_{dir}(\mathbf{x}_{i,dir}|\boldsymbol{\theta}_j)\} \tag{43}$$

and for the Two-layer model, because of its two latent variables, it read as

$$\log p_{\text{two-layer}}(\mathbf{x}_i, y_i, z_i) = \sum_{j}^{K_{loc}} y_{ij} \left(\log\{\pi_j f_{loc}(\mathbf{x}_{i,loc}|\boldsymbol{\theta}_j)\} + \sum_{k}^{K_{dir}} z_{ik} \log\{\omega_{jk} f_{dir}(\mathbf{x}_{i,dir}|\boldsymbol{\theta}_{jk})\} \right). \tag{44}$$

In addition, we also measure the training running time and total amount of iterations until convergence.

4.5 Initialization Methods

We initialize the posterior matrix and then perform a single M-step to obtain starting parameters $\gamma = \{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^{K}$. We consider four schemes:

1. **k-Means++ seeding** – Select K initial centroids using the k-means++ algorithm [9]. Each data point selected is the unique member of each cluster, and they will lead to the MLE parameter for each component distribution $\boldsymbol{\theta}_j$. The priors are set to $\pi_j = 1/K$.
2. **k-Means** – Run k-means with k-means++ initialization and set posteriors to the one-hot encodings of the resulting labels.
3. **Random from data** – Sample K observations uniformly at random to seed components (similar to k-means++ seeding but without distance-based selection) and set $\pi_j = 1/K$.
4. **Random** – Each member of the posterior probability matrix is randomly drawn from a uniform distribution.

5 Results

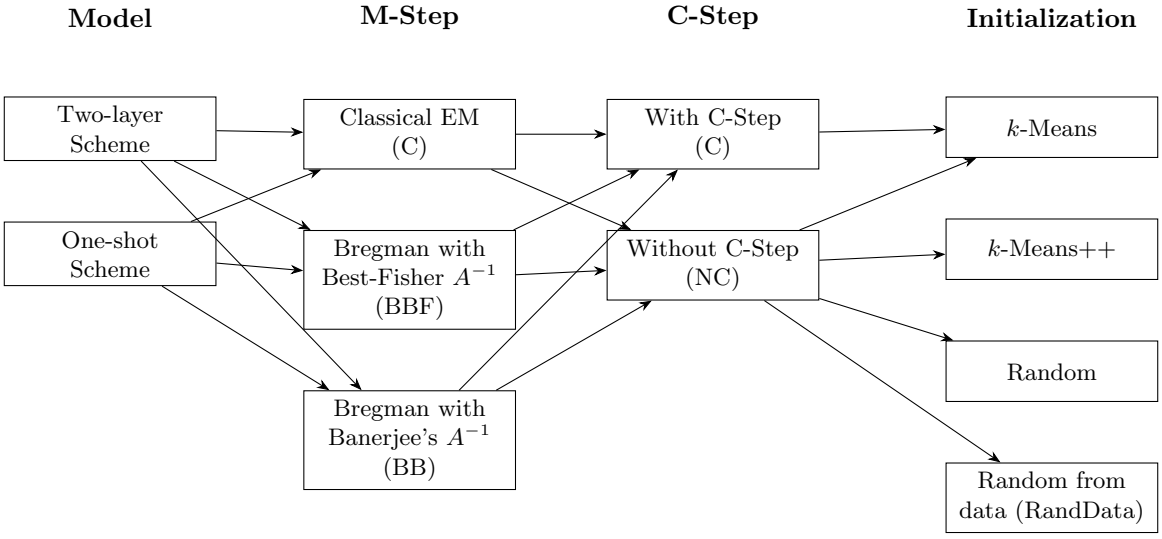


Figure 4: Combinations of model, M–step, C–step, and initialization compared in our experiments.

We evaluate the proposed clustering models on on–ball actions from the 64 matches of the 2018 FIFA World Cup, using event data provided by StatsBomb. We consider ten action types: clearances, corners, crosses, dribbles, free–kicks, goal–kicks, goalkeeper actions, passes, shots, and throw–ins. For each action type we fit both the hierarchical Two–layer model and the One–shot model under 15 different algorithmic configurations, given by the combinations of M–step variant, presence or absence of the C–step, and initialization scheme Figure 4).

When the C–step is included, we restrict the initialization to k-means only; for the other initialization schemes we observed numerical instabilities when followed by hard reclassification.

5.1 Model comparison

For each configuration we report four quantities: BIC, ICL, total running time, and the number of EM iterations until convergence. In all four metrics, smaller values are better. To compare the Two–layer and One–shot schemes we define, for each configuration,

$$D = M_{\text{Two-layer}} - M_{\text{One-shot}}, \tag{45}$$

where M denotes any of the four metrics. Thus $D > 0$ indicates that the One–shot model outperforms the Two–layer model (lower BIC/ICL, faster, or fewer iterations).

Across the 150 configurations considered, the One–shot model achieves lower BIC in 82% of the cases, lower ICL in 91%, shorter running time in 87%, and requires fewer iterations in 100% of the cases. When the Two–layer model does perform better, the advantage is typically modest. The joint distribution of the metric differences is shown in Figure 5, where points in the positive half–planes correspond to configurations where the One–shot scheme dominates.

5.2 M-Step Variants

We now compare the three M–step implementations: the classical EM M–step (C), the Bregman M–step using Banerjee's inverse A^{-1} (BB), and the Bregman M–step using the Best–Fisher inverse A^{-1} (BBF). Figure 6 and Figure 7 summarize the results for the One–shot and Two–layer models respectively.

Overall, the Bregman variants provide a small but consistent improvement in BIC and ICL relative to the classical

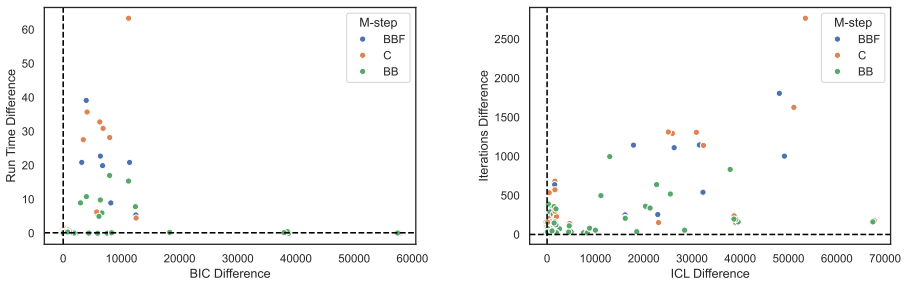


Figure 5: Model comparison between the One–shot and Two–layer schemes. Positive values on the axes indicate configurations where the One–shot model outperforms the Two–layer model.

M–step, with the effect more noticeable for the Two–layer model. The main differences, however, appear in computational efficiency: both Bregman implementations reduce running time and the number of EM iterations compared with the classical approach, with BB typically performing best among the two Bregman options. This confirms that replacing the numerical optimization for the von Mises parameters by the Bregman update yields tangible efficiency gains without hurting model fit at the level of BIC/ICL.

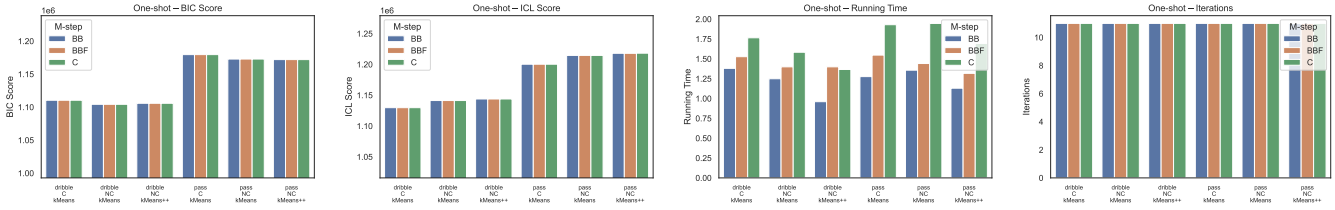


Figure 6: Comparison of M–step variants for the One–shot model across different action types and initialization schemes.

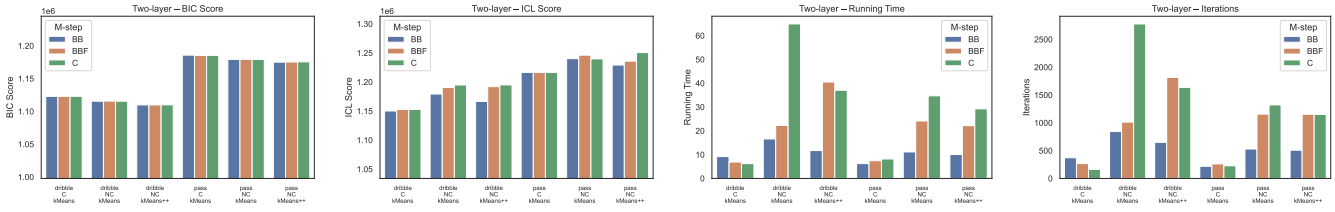


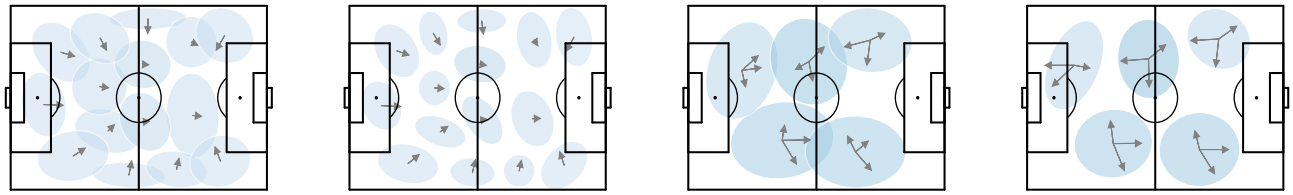
Figure 7: Comparison of M–step variants for the Two–layer model across different action types and initialization schemes.

5.3 Effects on C-Step

Then, we study the impact of adding a C–step (hard reclassification) inside each EM iteration. Quantitatively, the C–step generally leads to slightly lower ICL and clearer cluster separation, as expected from a classification–oriented criterion. Qualitatively, the resulting clusters are more compact and have sharper boundaries, as illustrated for passes in Figure 8: from left to right we show the One–shot model without and with C–step, followed by the Two–layer model without and with C–step.

We do not view the C–step variant as a definitive replacement for the standard soft-EM solution. In practice, the two perspectives are complementary: the soft assignments preserve uncertainty and smoother transitions between

neighboring patterns, while the C–step yields crisper zones. For exploratory analysis of soccer case studies it is therefore useful to inspect both versions side by side, as they highlight different but mutually reinforcing aspects of the underlying action structure.



(a) One–shot without C–step (b) One–shot with C–step (c) Two–layer without C–step (d) Two–layer with C–step

Figure 8: Effect of the C–step on the learned clusters for passes.

5.4 Initialization

Finally, we compare four initialization strategies: standard k-means, k-means++, purely random responsability components r_{ij}, and random selection of component centers from the data (RandData). The results are summarized in Figure 9 in terms of BIC, ICL, running time, and number of EM iterations.

Purely random initialization consistently performs worst across all metrics and exhibits high variability between runs, making it unsuitable in practice. Initializing from randomly chosen data points (RandData) occasionally discovers interesting local optima, but remains markedly unstable and typically underperforms the structured alternatives.

The main comparison is therefore between k-means and k-means++. Overall, their BIC and ICL values are very similar, but k-means shows more robust behavior when combined with the C–step, leading to more reliable convergence. Given this robustness and its low computational cost, we adopt k-means as our default initialization method for all subsequent experiments.

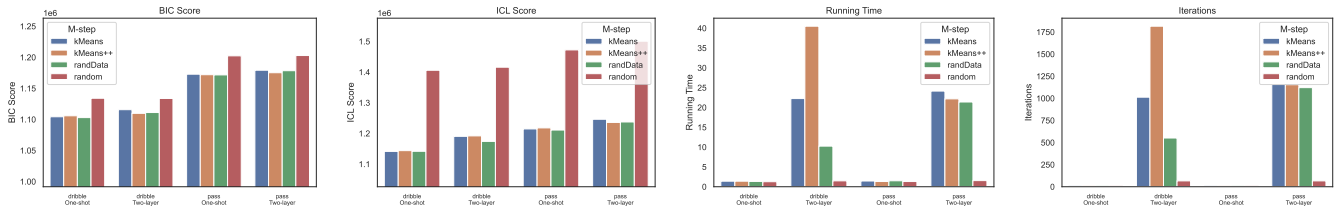


Figure 9: Comparison of initialization methods.

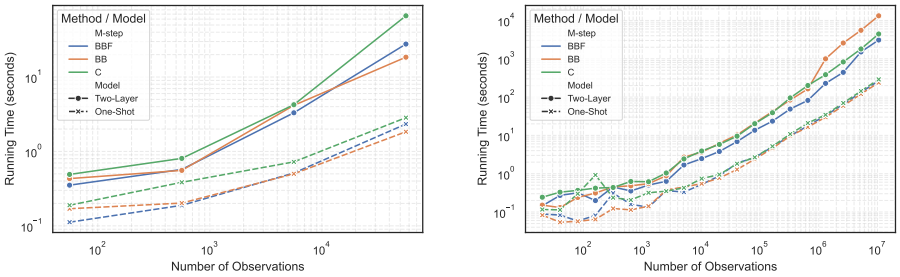
5.5 Time Complexity

We now investigate how the different model variants scale with the number of observations N. For both the real and synthetic experiments we fix the number of components and convergence tolerance, and we report wall–clock running time for the full EM procedure. Figure 10 shows the results on a log–log scale.

On the real World Cup data (Figure 10a), running time grows approximately linearly with N, in line with the $\mathcal{O}(NK)$ cost of each EM iteration. For any fixed M–step, the Two–layer model (solid lines) is consistently slower than the One–shot model (dashed lines), reflecting the additional inner mixture over directions. Among the M–steps, the classical update (C) is always the slowest option, while the two Bregman variants (BB and BBF) reduce running time by a noticeable margin.

The synthetic experiment (Figure 10b) explores a wider range of sample sizes, up to $N = 10^7$. The same qualitative picture emerges: (I) both models exhibit linear scaling in N; (II) the Two–layer scheme is roughly a constant factor slower than the One–shot scheme; and (III) Bregman-based M–steps are generally more efficient than the classical M–step. For large N, the BBF variant is the most stable and fastest of the three, whereas the classical M–step and the BB variant become increasingly expensive.

Overall, these experiments indicate that the proposed One–shot model combined with a Bregman M–step (in particular BBF) provides the most favorable trade–off between modeling flexibility and computational cost, especially at the dataset sizes typical of modern event data.



(a) Real dataset

(b) Synthetic dataset

Figure 10: Empirical time complexity.

6 Conclusion

In this work we addressed the unsupervised discovery of spatio-directional on-ball action structure in soccer by clustering its start location and direction. Building on the exponential family and Bregman divergence framework, we reformulated EM in expectation-parameter (dual) space so that M-steps become Bregman centroid updates, yielding closed-form updates for any exponential family distribution thus avoiding numerical optimization in the case of von Mises precision parameter via accurate analytic inverses of the mean resultant length. We instantiated the framework in two strategies, a Two-layer hierarchical mixture and a One-shot joint simultaneous mixture, and benchmarked classical vs. Bregman M-steps, optional hard reclassification (C-step), and multiple initializations on StatsBomb event data from the 64 matches of the 2018 FIFA World Cup across ten action types.

Empirically, the One-shot scheme dominated the Two-layer alternative across most configurations (lower BIC in 82% of cases, lower ICL in 91%, faster in 87%, and fewer iterations in 100%), and Bregman M-steps delivered tangible efficiency gains without hurting fit. From a practical standpoint, our experiments suggest a clear default recipe for event datasets: use the One-shot model with a Bregman M-step (particularly the BBF variant for stability at large N), initialize with k-means for robust convergence, and treat the C-step as a complementary lens rather than a replacement — soft assignments preserve uncertainty while C-step hardening yields crisper, more compact zones and tends to slightly improve classification-oriented criteria like ICL. More broadly, the linear runtime scaling observed across both real and synthetic settings supports the feasibility of these mixtures for large collections of actions.

Key limitations remain in the modeling assumptions — most notably the conditional-independence structure used to relate location and direction in the hierarchical construction — and in the restricted feature set, which omits temporal and contextual information about how actions unfold. Natural next steps are therefore twofold: on the methodological side, relax these assumptions and extend the action representation to capture richer latent structure; on the applied side, translate the empirical findings into a practitioner-facing "playbook" that maps modeling choices (scheme, M-step variant, C-step, initialization) to concrete use cases and decision-making needs for analysts and coaching staff.

References

[1] T. Decroos, M. V. Roy, and J. Davis, "Soccermix: Representing soccer actions with mixture models," 2021. Available from KU Leuven, Belgium.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[4] S.-i. Amari, *Information Geometry and Its Applications*. Tokyo: Springer, 2016.

[5] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises–fisher distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[6] D. J. Best and N. I. Fisher, "The bias of the maximum likelihood estimators of the von mises–fisher concentration parameters," *Communications in Statistics – Simulation and Computation*, vol. 10, no. 5, pp. 493–502, 1981.

[7] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Computational Statistics & Data Analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[8] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

[9] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, (New Orleans, Louisiana, USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, Jan. 2007.

7 Annex

7.1 Clusters Visualization

The following plots are the clusters obtained by the two-layer and one-shot models for the different action datasets. Rows are M-steps approaches (BBF, C, BB in that order), and columns are initialization methods (k-Means, k-Means++, RandData, and Random)

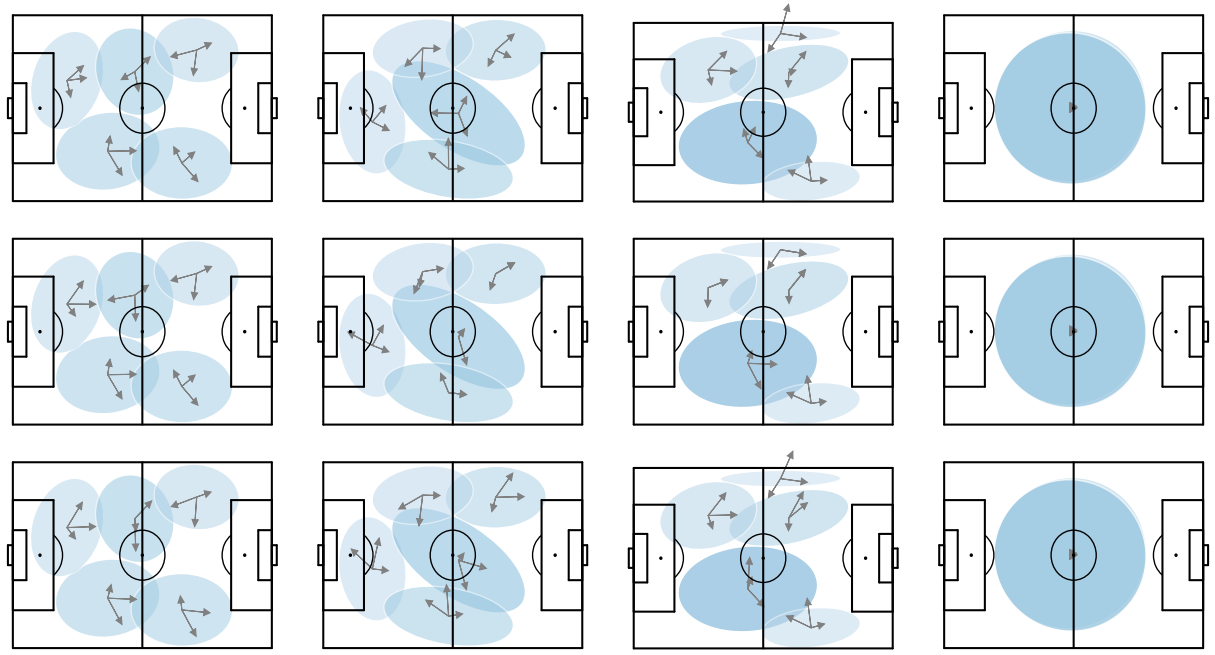


Figure 11: Two-layer Model — Passes

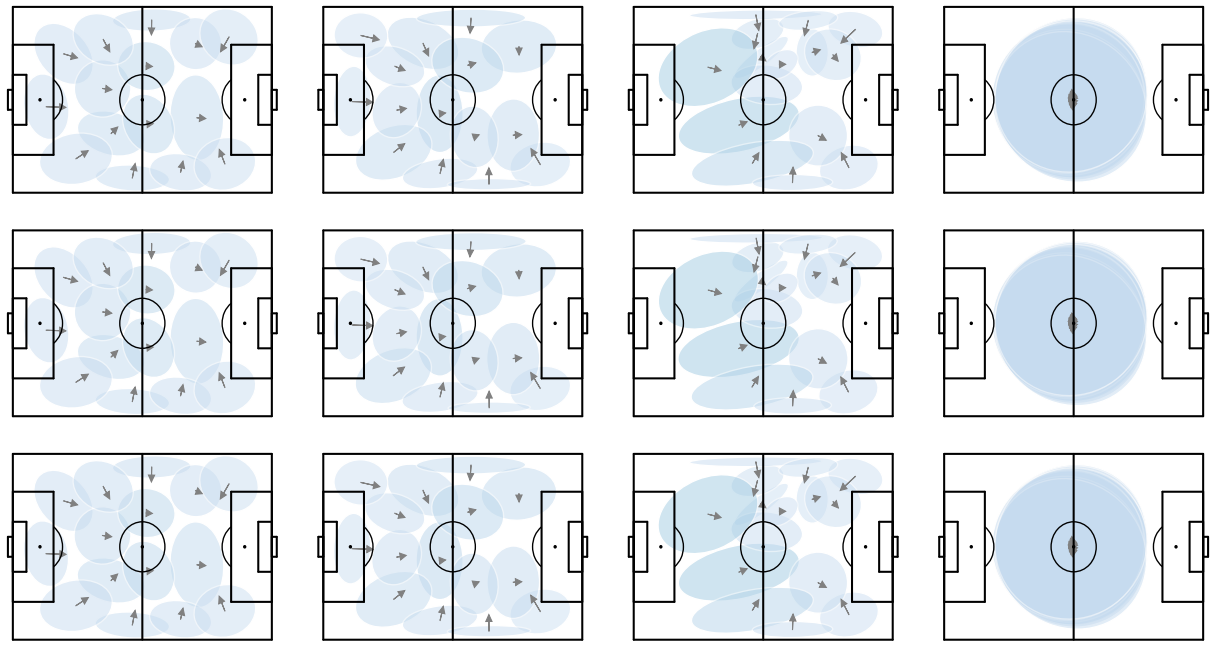


Figure 12: One-shot Model — Passes

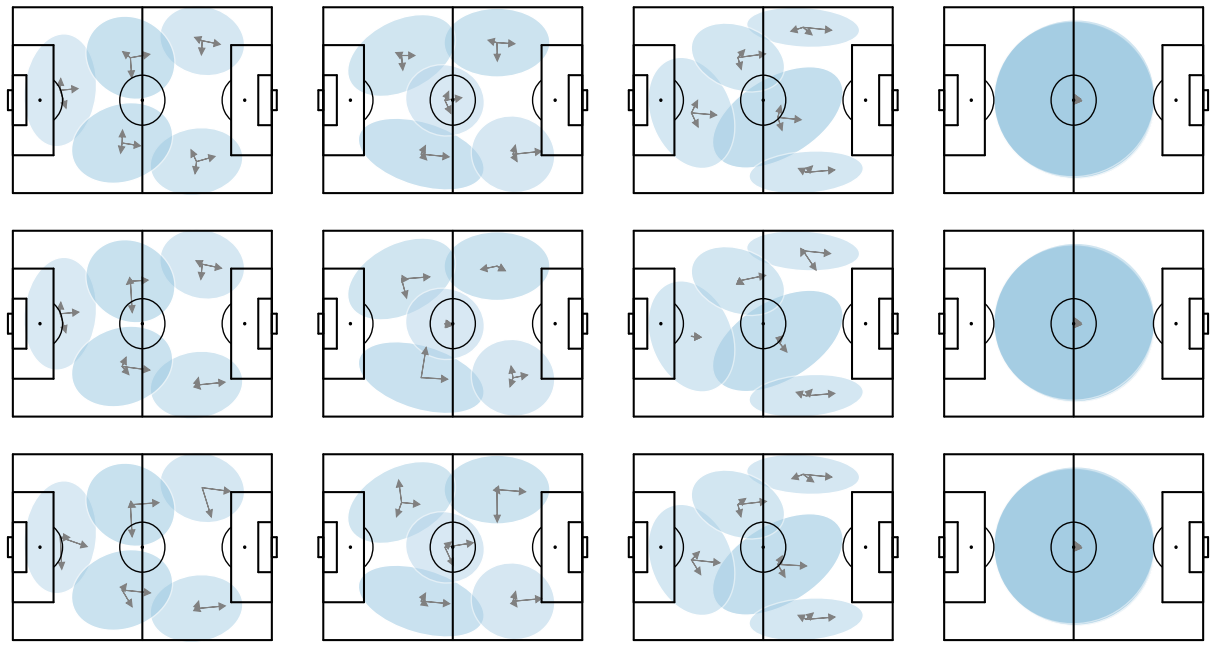


Figure 13: Two-layer Model — Dribbles

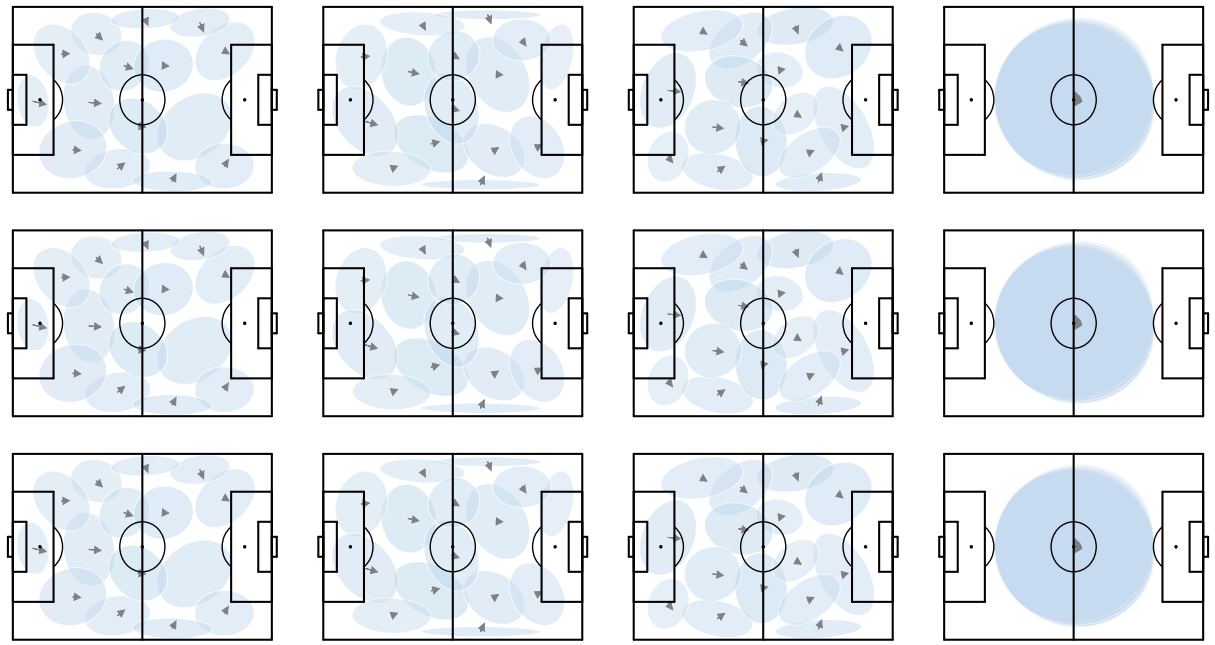


Figure 14: One-shot Model — Dribbles

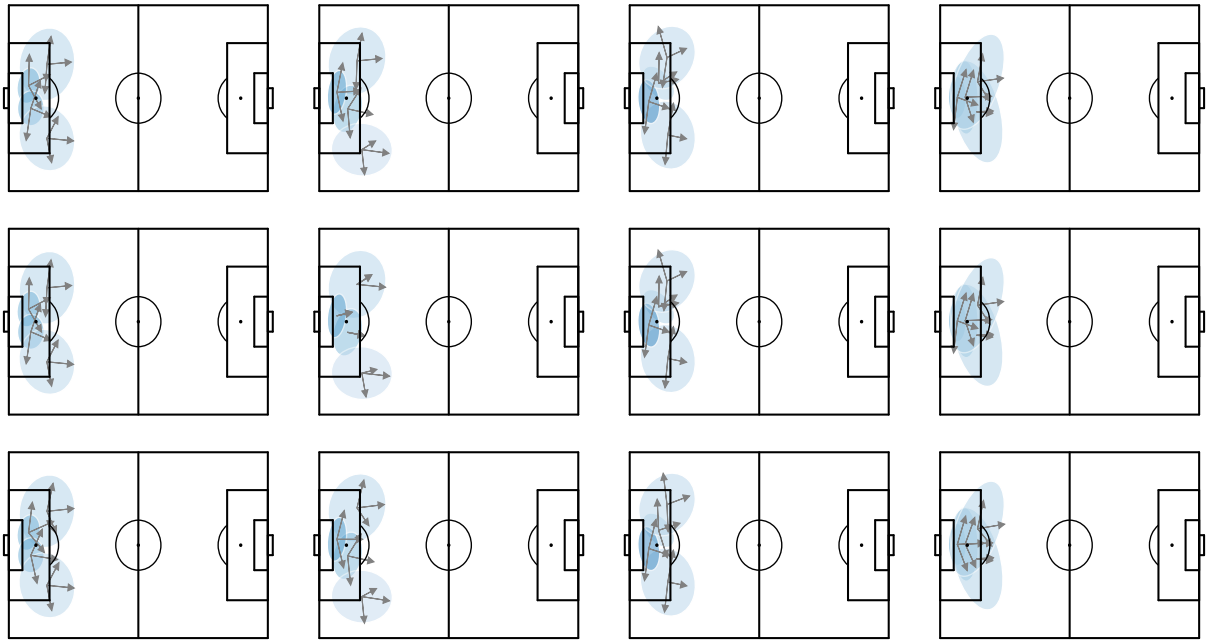


Figure 15: Two-layer Model — Clearances

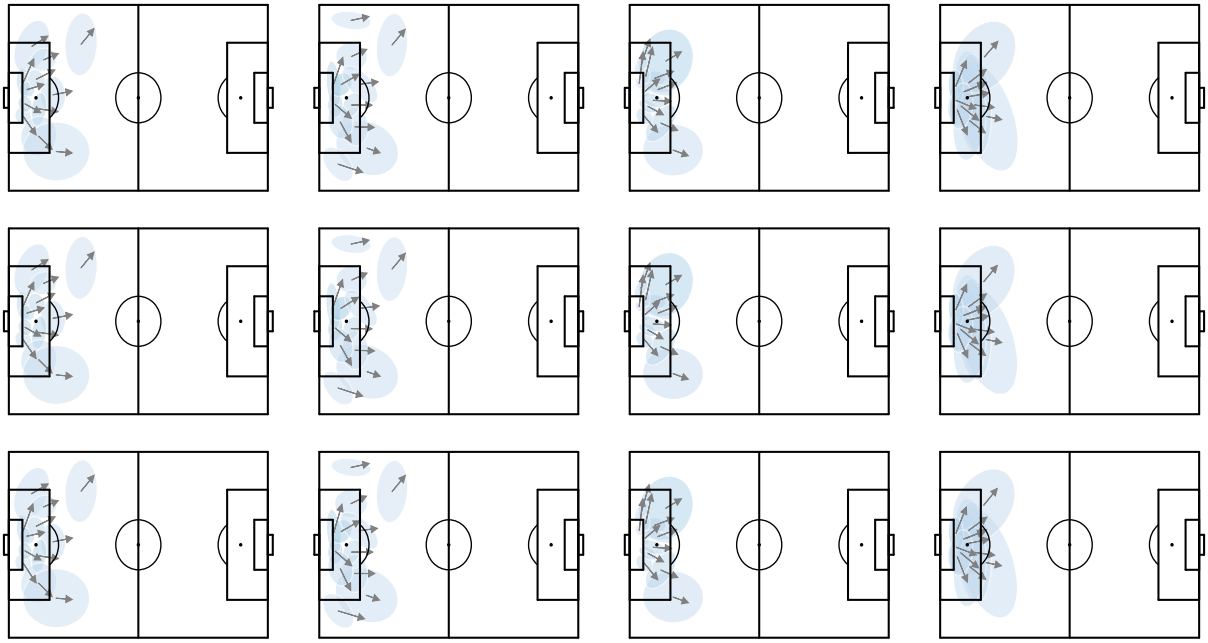


Figure 16: One-shot Model — Clearances

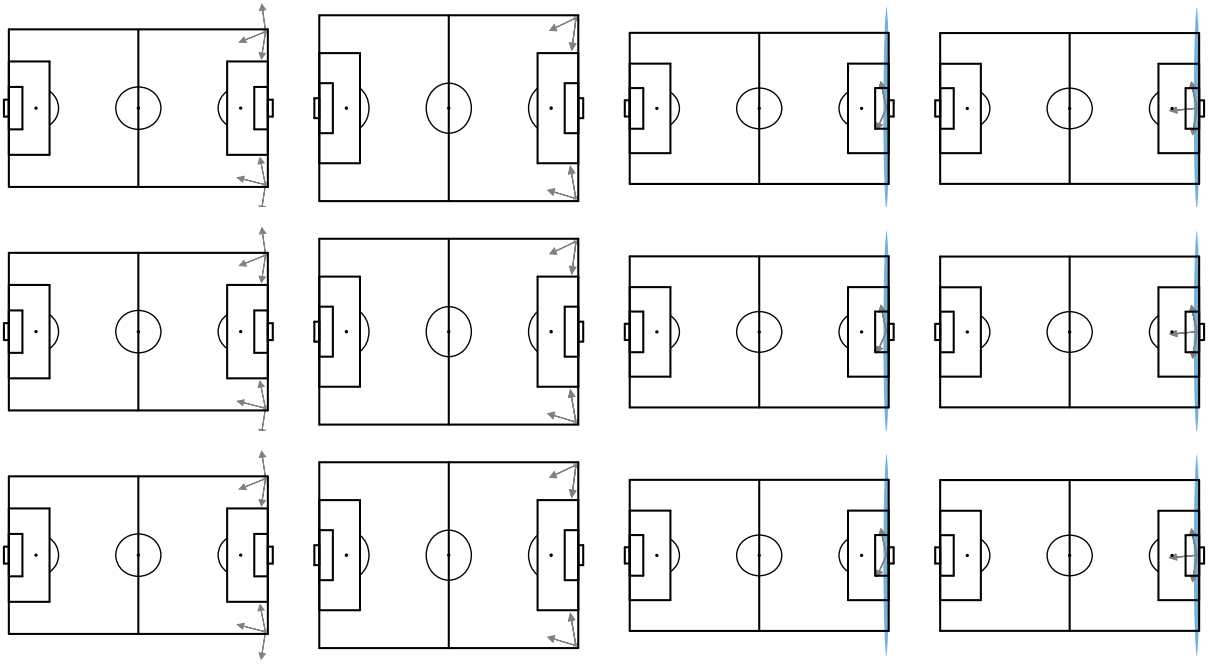


Figure 17: Two-layer Model — Corners

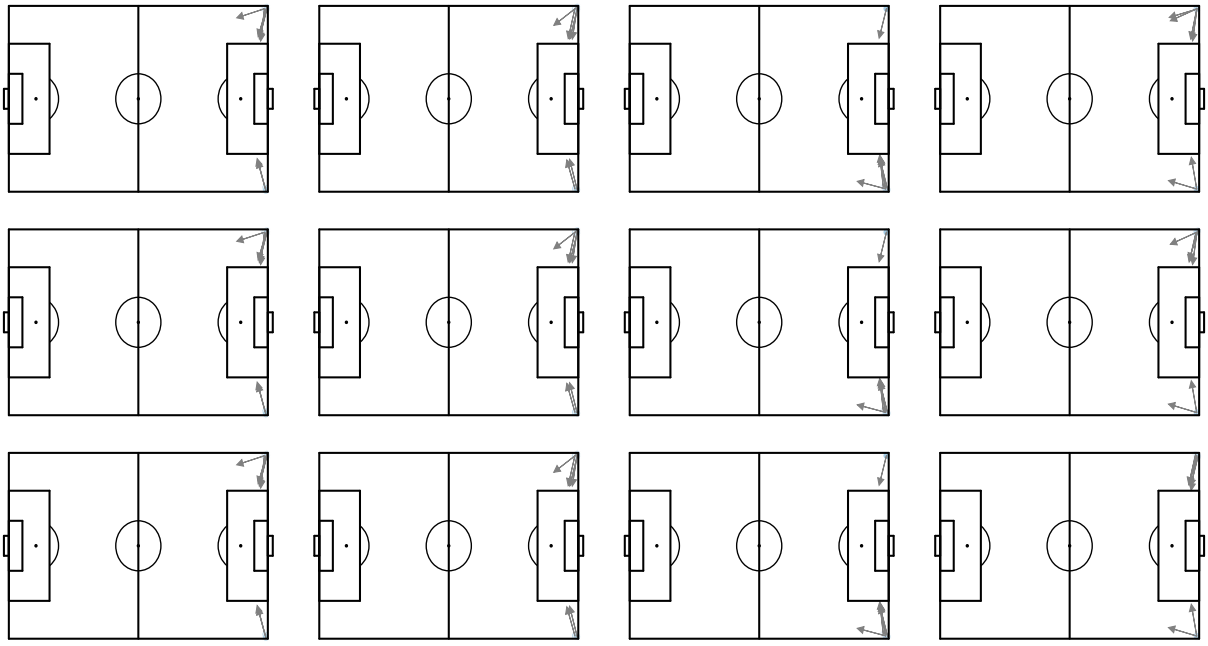


Figure 18: One-shot Model — Corners

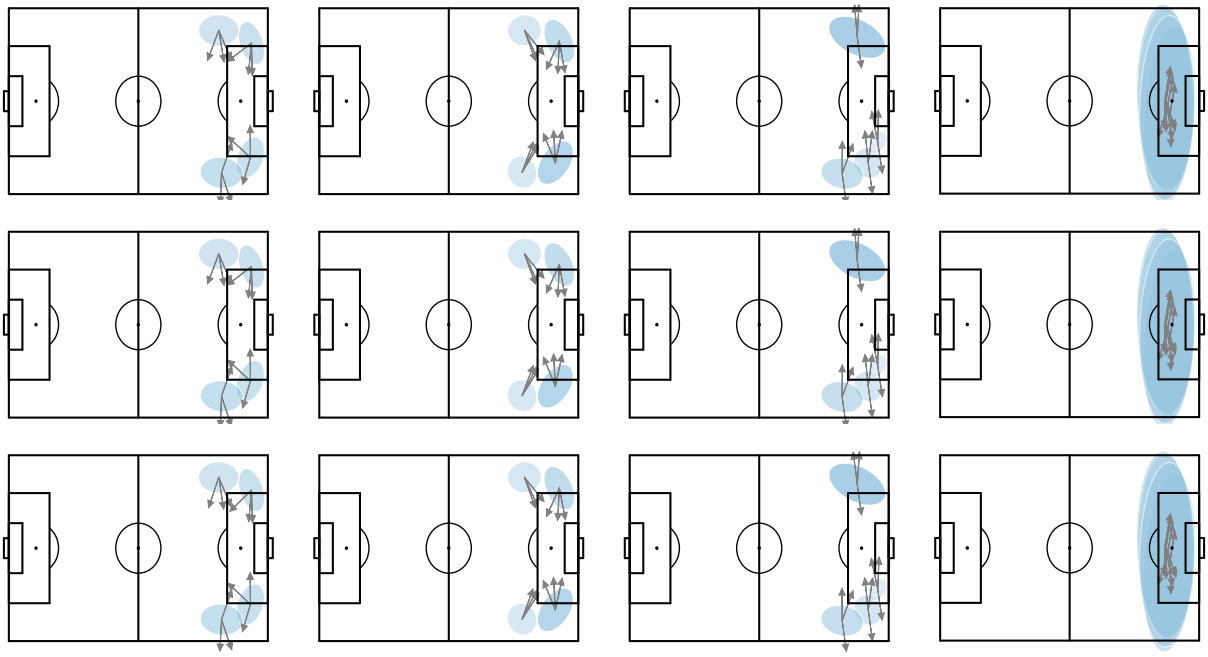


Figure 19: Two-layer Model — Crosses

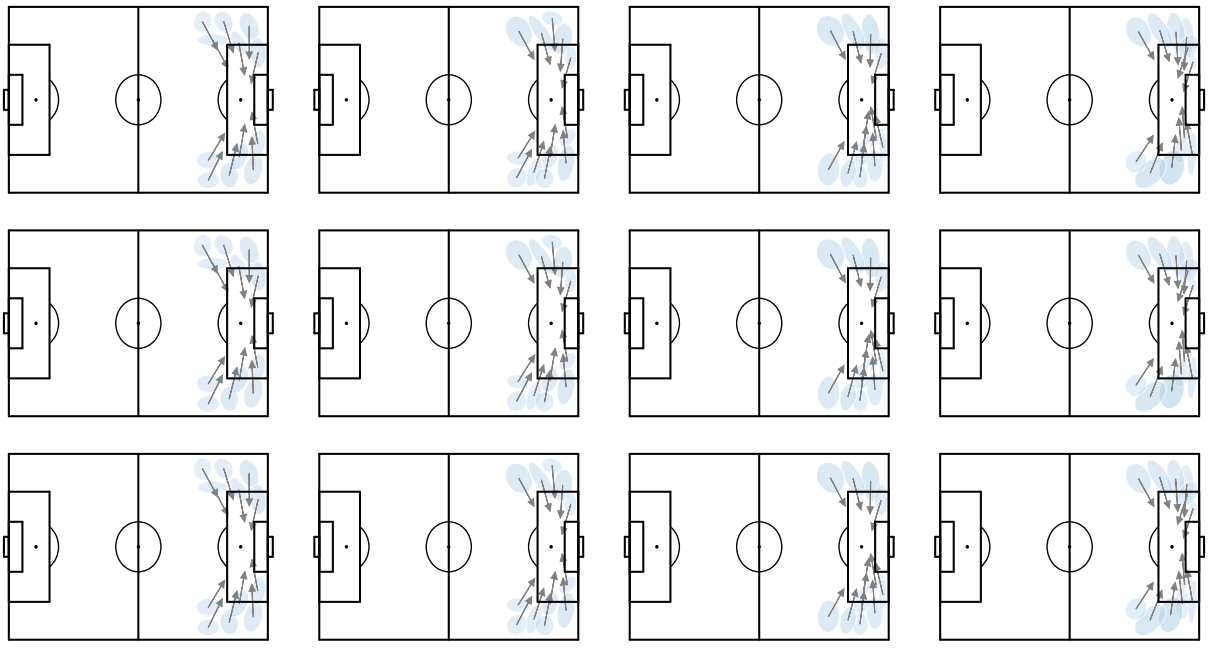


Figure 20: One-shot Model — Crosses

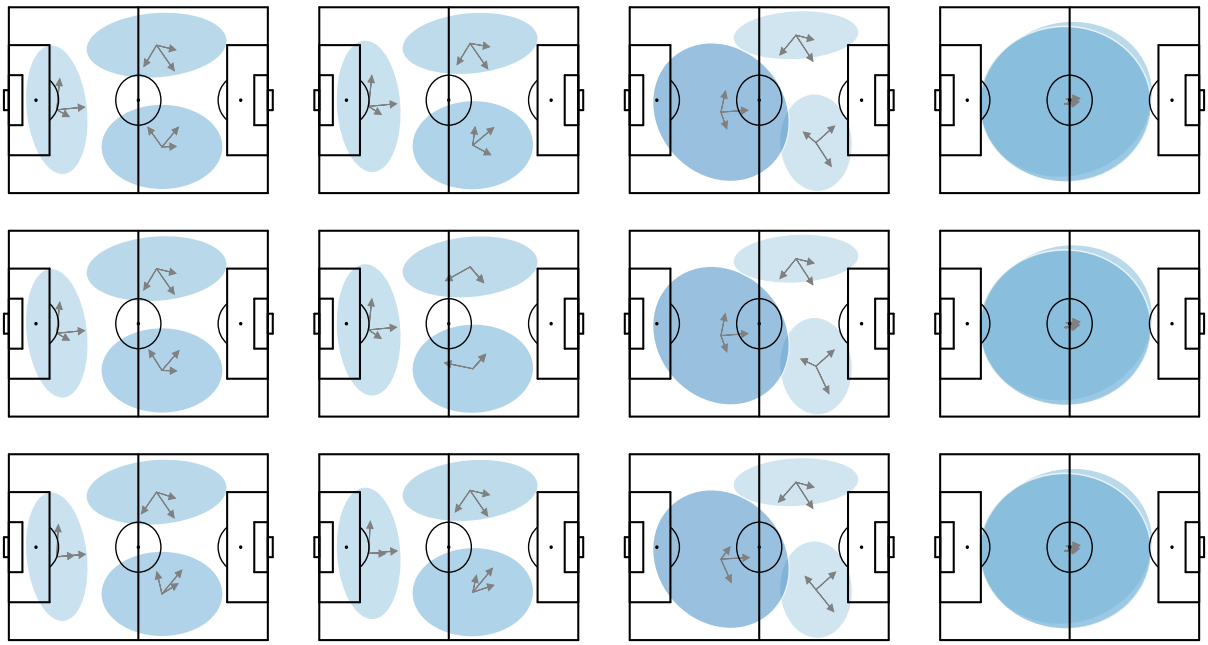


Figure 21: Two-layer Model — Free Kicks

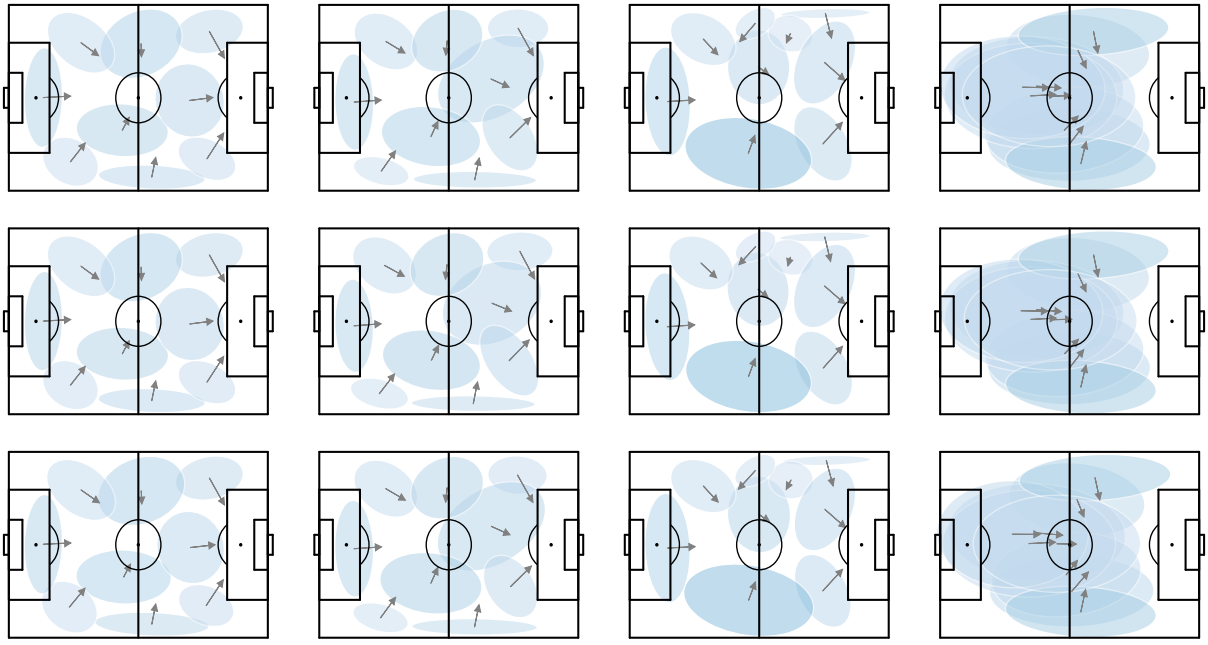


Figure 22: One-shot Model — Free Kicks

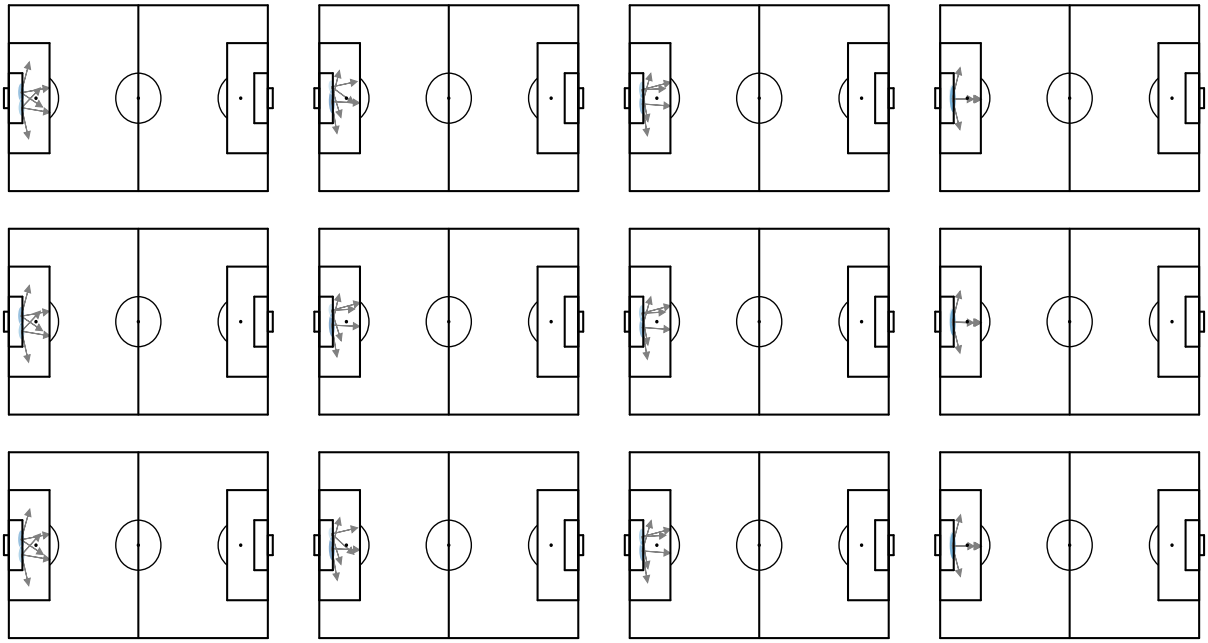


Figure 23: Two-layer Model — Goal Kicks

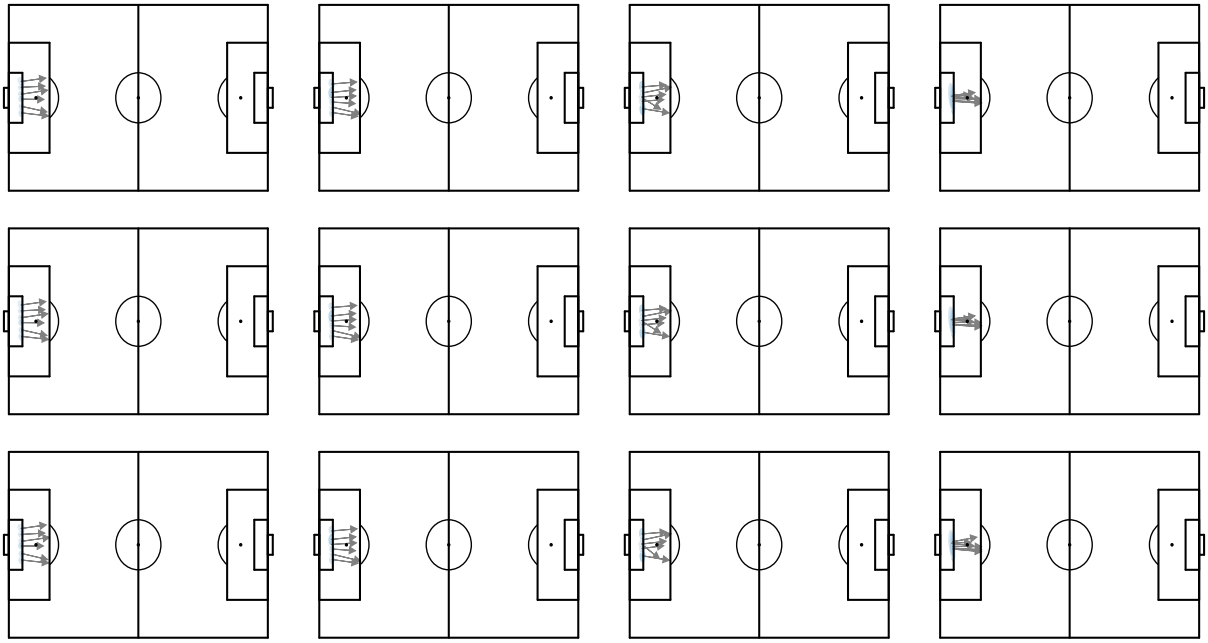


Figure 24: One-shot Model — Goal Kicks

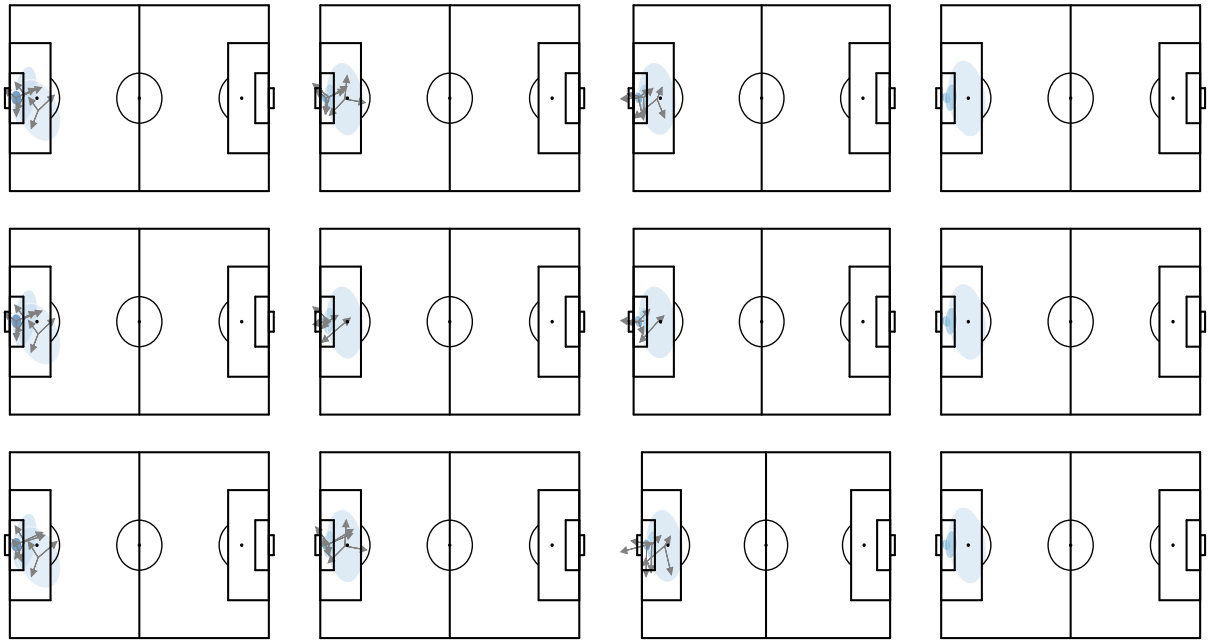


Figure 25: Two-layer Model — Keeper Actions

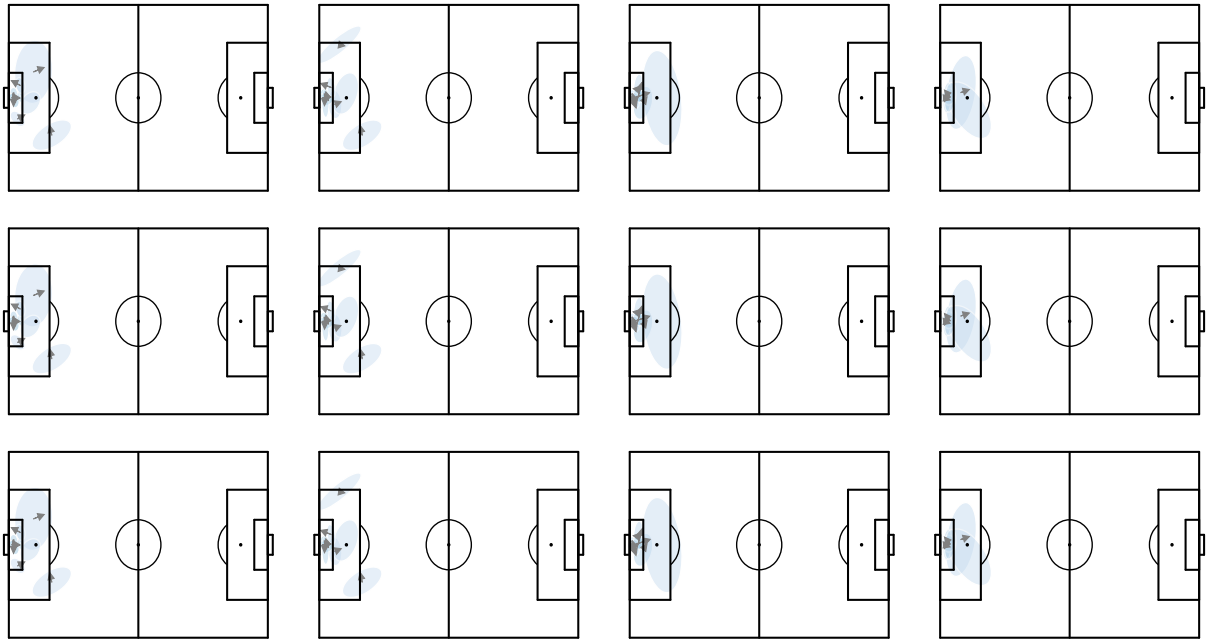


Figure 26: One-shot Model — Keeper Actions

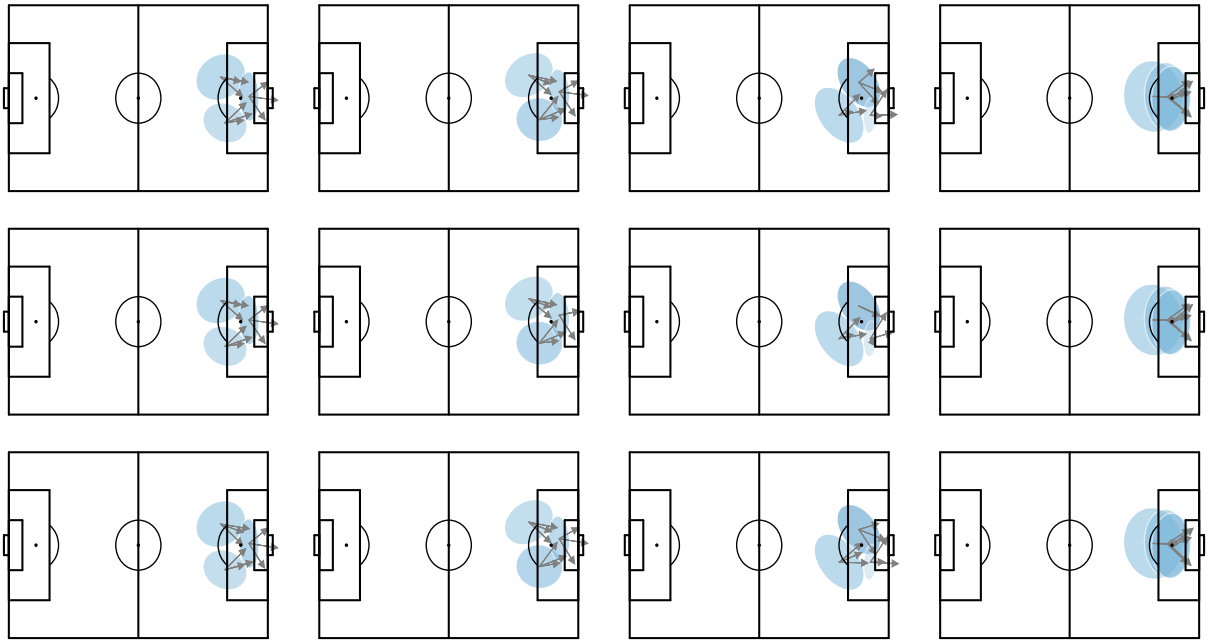


Figure 27: Two-layer Model — Shots

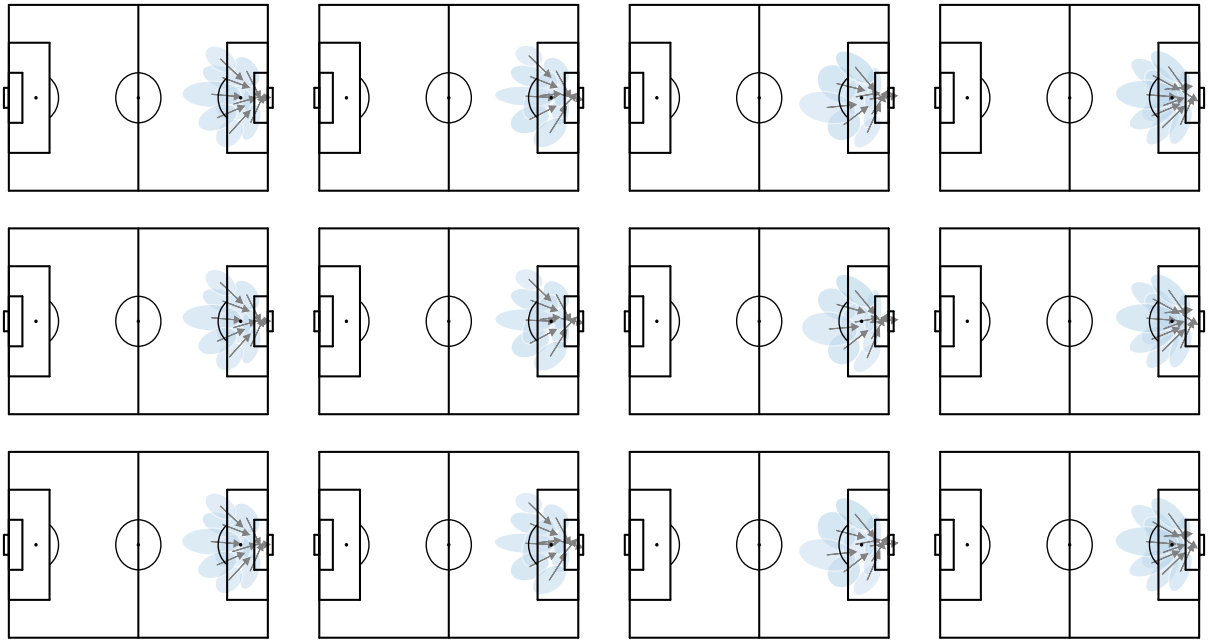


Figure 28: One-shot Model — Shots

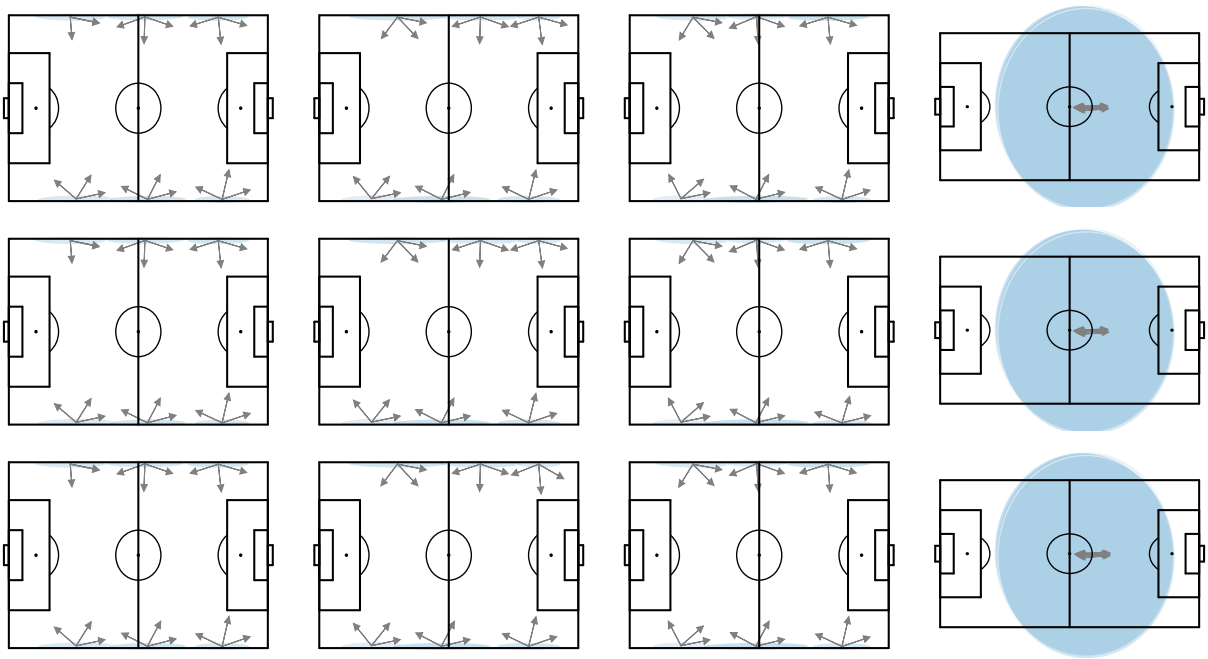


Figure 29: Two-layer Model — Throw-ins

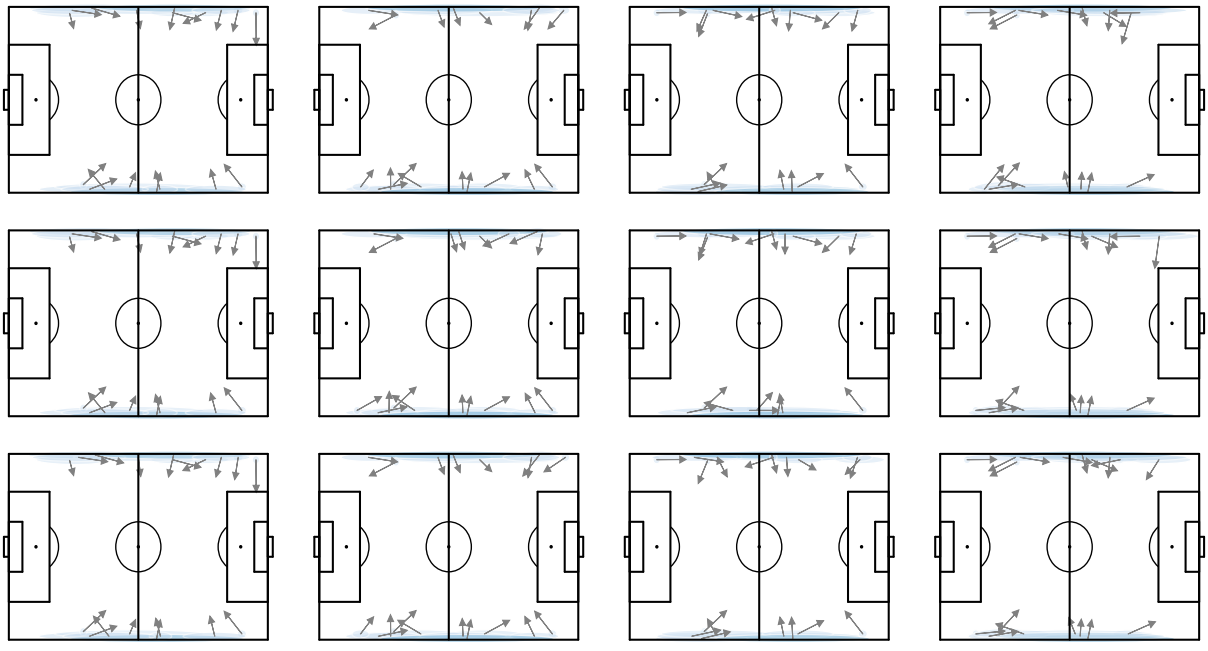


Figure 30: One-shot Model — Throw-ins