Zone-Based Indoor Localization from Incomplete RSSI Fingerprints via Missing Data Marginalization

Leon Winheim, **Daniel Frisch** and **Uwe D. Hanebeck**
Intelligent Sensor-Actuator-Systems Laboratory (ISAS)
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
{leon.winheim, daniel.frisch, uwe.hanebeck}@{ieee.org, kit.edu}

***Abstract*—We present a novel imputation-free approach for probabilistic zone classification in scenarios with highly uncertain, partially incomplete Received Signal Strength Indicator (RSSI) fingerprints from Bluetooth Low Energy (BLE) beacons. RSSI measurements from multiple beacons, received at the current time step, form a so-called fingerprint that is used to attribute a discrete zone label indicating where the receiver currently resides. A challenge in this prediction is that single beacons may not be observable at every point in time, which leads to incomplete measurements. We perform classification based on Gaussian mixture density estimation, with the special case of axis-aligned mixture models. This can capture arbitrary joint densities and is not merely a naive Bayes classifier. At the same time, this enables us to efficiently solve the problem of missing data by marginalization instead of imputation. We derive a Maximum Likelihood parameter estimator for learning and an online Maximum a-posteriori (MAP) classifier for localization, all grounded in the Bayesian framework. Zone estimation at runtime is very fast, because evaluating Gaussians with diagonal covariances does not involve cubic-runtime operations as for general Gaussians. We evaluate our approach on a real-world dataset with 210,000 fingerprints from 21 beacons and demonstrate its performance (F_1 score $> 97\,\%$) while being computationally inexpensive.**

***Index Terms*—Indoor Localization, Positioning, Navigation, RSSI, Bluetooth Low Energy (BLE), Missing Values, Incomplete Data, Imputation, Marginalization, Supervised Learning, Gaussian Mixture Model, Density Estimation, Zone-Based Localization, Probabilistic Fingerprinting, Maximum Likelihood Estimation, IoT, Smart Building.**

I. Introduction

A. Overview

In many applications, such as asset tracking, navigation, and location-based services, accurate indoor localization is crucial. Indoor localization is generally challenging due to the lack of GNSS signals. Available technologies for highly precise indoor localization require expensive sensors, such as LiDAR. Camera-based optical tracking systems require extensive and error-prone infrastructure, such as paving the floor with ArUco or AprilTag markers. Ultra-Wideband (UWB) systems pollute the frequency spectrum and are expensive as they need accuracy in the nanosecond range.

This work was funded by the Federal Ministry for Economic Affairs and Climate Action as part of a project of the Central Innovation Programme for SMEs (ZIM).

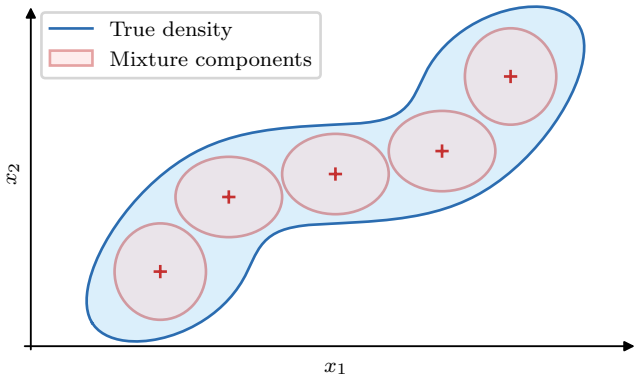


Fig. 1: Non-Gaussian density (blue), approximated by GM with axis-aligned components (red ellipses).

As an alternative technology, Bluetooth beacons are a cost-effective and widely available solution for indoor localization. They can easily be deployed in various environments and provide Received Signal Strength Indicator (RSSI) data that can be used for estimating the location of devices [1]. However, RSSI data is highly uncertain and often incomplete due to signal interference and attenuation, multipath propagation, and device limitations.

In addition, the computational resources can be a limiting factor, e.g., if computation happens on mobile devices with limited embedded processing capabilities and power constraints. Therefore, efficient algorithms are needed to process RSSI data during training, but especially during (online) prediction.

B. Zone-Based Localization

In this work, we focus on zone-based indoor localization, which aims to classify the location of devices into predefined zones rather than providing precise coordinates. This approach can be faster and more practical, especially when only zone-level information is required. Zone-based localization is sufficient for many applications, such as room-level localization or asset tracking within specific areas.

The RSSI measurements from multiple beacons at a given time step form a so-called *fingerprint*, which serves as input to a classifier that predicts the corresponding zone. Technically, the task can be solved by any classification method, examples

being support-vector machines (SVMs), random forests, or Gaussian mixture (GM) models.

C. Missing Data

A key challenge in RSSI-based fingerprinting is that individual beacons may not be observable at every time step, resulting in partially missing measurements. We treat the special case of data that is "missing at random" [2]. Any classification method employed for this task requires a strategy to cope with such incomplete inputs. Common approaches include discarding incomplete measurements, which can be wasteful [3], or imputing surrogate values, which can introduce bias and be costly [2]. Gaussian and GM models offer an alternative via closed-form marginalization, i.e., integrating out the missing dimensions rather than guessing their values. However, for full-covariance models, evaluating the exponents of the Gaussian components requires matrix inversions for each specific pattern of missing dimensions, which can be expensive when many beacons are involved.

D. Contribution

We propose using GMs with axis-aligned components (see Fig. 1 for a visualization) for zone-based RSSI classification. This leads to the following advantages:

- Missing data is handled formally correct by marginalization, which for axis-aligned Gaussians reduces to simply dropping factors from a product, and therefore no imputation or matrix inversion is needed.
- The method admits closed-form maximum likelihood (ML) parameter estimation of the Gaussian density parameters, even in the presence of missing values.
- Unlike a naive Bayes classifier, the placement of multiple axis-aligned components can capture complex, multimodal distributions.
- For N_B beacons, the per-prediction cost scales with $\mathcal{O}(N_B)$ instead of $\mathcal{O}(N_B^3)$ for full-covariance models, making it suitable for resource-constrained embedded devices.

We evaluate the method on a real-world dataset and compare it to other state-of-the-art methods, reaching comparable performance while maintaining low computational complexity.

E. Structure of the Paper

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed method. Section IV contains the evaluation, and Section V concludes the paper.

II. Related Work

RSSI-based localization methods can be categorized along multiple dimensions. One of these is the network protocol used, with examples being Bluetooth and Wi-Fi [4]. In general, any protocol is feasible where RSSI values can be extracted. Another dimension is the type of the result. It can either be a regression-type prediction of specific coordinates, e.g., (x, y) in two dimensions, or a classification-type output, pointing to one out of a set of predefined zones [4].

A. Classic Machine Learning Methods

The focus of this work is on classification for discrete zones. Classic methods for this task are, for example, SVMs, k-nearest neighbor classifiers, or random forests. For a survey on these classic methods, we refer to [4], [5]. Deep learning approaches have also been applied to RSSI-based fingerprinting, e.g., [6], but they are typically data-hungry, act as black-box models, and do not offer a principled mechanism for handling missing input dimensions. These methods are often used in combination with imputation strategies in case incomplete data is not discarded.

B. Missing Data Strategies

The problem of partially missing data is a fundamental challenge in RSSI-based localization [2] and can be tackled by the following approaches.

Discard: The simplest approach is to discard incomplete fingerprints entirely (listwise deletion), but this can be wasteful and may not be viable if missing data is frequent [3].

Impute: More sophisticated approaches include imputation methods which substitute a surrogate value for each missing entry, e.g., a mean value, a k-nearest neighbor estimate, or an iteratively refined estimate [2]. However, any imputed value is based on specific assumptions and can induce unwanted bias.

Marginalize: A third category is model-based marginalization, where no single value is guessed but the distribution over the missing dimensions is integrated out. This is the statistically principled approach. For most machine learning models, this would require expensive numerical integration, but Gaussian and GM models have a closed-form analytic solution to this: simply remove the respective rows and columns from the covariance matrix and the mean vector. However, evaluating the Mahalanobis distance $(\underline{x}-\underline{\mu})^\top \mathbf{C}^{-1}(\underline{x}-\underline{\mu})$ in the Gaussian exponent requires either i) solving the system of linear equations $\mathbf{C}\underline{a} = (\underline{x} - \underline{\mu})$ for $\underline{a}$ or ii) computing and caching the $\mathbf{C}^{-1}$ matrices for all 2^{N_B} possible patterns of missing measurements, with subsequent vector-matrix multiplication $\underline{a} = \mathbf{C}^{-1}(\underline{x} - \underline{\mu})$. In both cases, the operations have roughly $\mathcal{O}(N_B^3)$ complexity at runtime, where (ii) has a better constant but exponential storage requirement. Therefore, we propose using diagonal covariances, yielding a complexity of $\mathcal{O}(N_B)$.

C. Gaussian Mixture (GM) Models for Localization

Alternatives to the classic machine learning approaches are provided by the mentioned probabilistic classification methods based on Gaussian densities and mixtures thereof. In [7], a naive Bayes classifier was implemented, where the GM model consists of independent per-beacon densities and therefore neglects correlations and other dependencies of the RSSI between beacons. Multiple publications treat the case of multivariate GM models with full covariance matrices, e.g., [8], [9], [10], [11]. These full-covariance models are expressive, but handling missing data via marginalization requires inverting a different submatrix for every observed index set, leading to a per-prediction cost of $\mathcal{O}(N_B^3)$ with N_B beacons.

For fitting GM models, Expectation–Maximization (EM) is typically applied to infer the weights, means, and covariance

matrices. After its first introduction in [12], the popular paper by Ghahramani and Jordan [13] demonstrated the inherent capability of the EM algorithm to handle missing data by marginalization. Most modern publications refer to this seminal paper as their base strategy to handle missing data, e.g., [3], [14].

Simplifications using axis-aligned GM models appear in [15] and in [16] without derivation. In the axis-aligned case, marginalization over missing dimensions reduces to simply dropping factors from a product, avoiding matrix inversions entirely and reducing the per-prediction cost to $\mathcal{O}(N_B)$. This is a big advantage of our proposed approach. At the same time, the placement of multiple axis-aligned components can still capture complex, multi-modal distributions—unlike a naive Bayes classifier that uses one univariate density per measurement dimension, i.e., per Bluetooth beacon in our case, and is not able to exploit interdependencies of RSSIs from different beacons. A visualization of the proposed strategy is given in Fig. 1. To the best of our knowledge, axis-aligned GM models with closed-form ML parameter estimation under missing data have not yet been applied to zone-based RSSI classification.

III. Proposed Method

A. Problem Description

We have a set of N_Z zones of interest, denoted as $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{N_Z}\}$, and a set of N_B Bluetooth beacons $\{B_1, B_2, \ldots, B_{N_B}\}$. Each beacon B_b broadcasts its unique identifier and the corresponding RSSI data is collected by devices within the environment that want to predict the zone they are in. The collected RSSI data is represented as a fingerprint vector $\underline{x} = [x[1], x[2], \ldots, x[N_B]]^\top$, where $x[b]$ is the RSSI measurement from beacon B_b. For each beacon whose broadcast has not been received in the respective time window, the fingerprint vector $\underline{x}$ contains a missing value. For each fingerprint, we describe the index set of available RSSI measurements with $\mathcal{I} \subseteq \{1, 2, \ldots, N_B\}$. For each zone we have a dataset of L recorded fingerprints $\mathcal{D} = \{(\underline{x}_i, \mathcal{I}_i)\}_{i=1}^{L}$. The goal is to estimate the probability density function (PDF) of the RSSI data for each zone Z_z, denoted as $f(\underline{x} \mid Z_z) = f(\underline{x} \mid \mathcal{D}_z)$, and use it for zone classification. $\mathcal{D}_z$ denotes the set of measurements corresponding to a specific zone index z.

B. Gaussian Estimation

We start modeling with an axis-aligned Gaussian density with mean vector $\underline{\mu}$ and variance vector $\underline{C}$

$$f(\underline{x} \mid \underline{\mu}, \underline{C}) = \prod_{b=1}^{N_B} \mathcal{N}(x[b] \mid \mu[b], C[b]) \ . \tag{1}$$

For missing values, instead of imputation, we marginalize out the missing dimensions

$$f_{\mathcal{I}}(\underline{x} \mid \underline{\mu}, \underline{C}) = \int_{x[j]} f(\underline{x} \mid \underline{\mu}, \underline{C}) \, \mathrm{d}x[j] \ , \tag{2}$$

$$\forall j \in \{1, \ldots, N_B\} \setminus \mathcal{I} \ , \tag{3}$$

which results in a Gaussian density $f_{\mathcal{I}}$ with reduced dimension

$$f_{\mathcal{I}}(\underline{x} \mid \underline{\mu}, \underline{C}) = \prod_{b \in \mathcal{I}} \mathcal{N}(x[b] \mid \mu[b], C[b]) \ . \tag{4}$$

The likelihood over the entire dataset is then given by the product of the likelihoods using the individual samples $\underline{x}_i$

$$f(\mathcal{D} \mid \underline{\mu}, \underline{C}) = \prod_{i=1}^{L} f_{\mathcal{I}_i}(\underline{x}_i \mid \underline{\mu}, \underline{C}) \tag{5}$$

$$= \prod_{i=1}^{L} \prod_{b \in \mathcal{I}_i} \mathcal{N}(x_i[b] \mid \mu[b], C[b]) \ . \tag{6}$$

The negative log-likelihood is

$$-\log f(\mathcal{D} \mid \underline{\mu}, \underline{C}) \tag{7}$$

$$= \frac{1}{2} \sum_{i=1}^{L} \sum_{b \in \mathcal{I}_i} \left(\log(2\pi C[b]) + \frac{(x_i[b] - \mu[b])^2}{C[b]} \right) \tag{8}$$

$$= \frac{1}{2} \sum_{b=1}^{N_B} \#(i : b \in \mathcal{I}_i) \cdot \log(2\pi C[b]) \tag{9}$$

$$+ \frac{1}{C[b]} \sum_{i : b \in \mathcal{I}_i} (x_i[b] - \mu[b])^2 \ ,$$

where $\#(i : b \in \mathcal{I}_i)$ is the number of samples in $\mathcal{D}$ with available RSSI measurement from beacon B_b. The ML estimate of the density parameters $\underline{\mu}$ and $\underline{C}$

$$\hat{\underline{\mu}}, \hat{\underline{C}} = \arg\min_{\underline{\mu}, \underline{C}} \big\{ -\log f(\mathcal{D} \mid \underline{\mu}, \underline{C}) \big\} \tag{10}$$

can be obtained by setting the gradient of the negative log-likelihood with respect to the parameters to zero. The ML estimate of the mean is given by

$$0 = \frac{\partial}{\partial \mu[b]} \big(-\log f(\mathcal{D} \mid \underline{\mu}, \underline{C}) \big) \ , \tag{11}$$

$$0 = \sum_{i : b \in \mathcal{I}_i} (x_i[b] - \mu[b]) \tag{12}$$

$$\Rightarrow \hat{\mu}[b] = \frac{1}{\#(i : b \in \mathcal{I}_i)} \sum_{i : b \in \mathcal{I}_i} x_i[b] \ . \tag{13}$$

The ML estimate for the variance is given by

$$0 = \frac{\partial}{\partial C[b]} \big(-\log f(\mathcal{D} \mid \underline{\mu}, \underline{C}) \big) \ , \tag{14}$$

$$0 = \#(i : b \in \mathcal{I}_i) \cdot C[b]^{-1} - C[b]^{-2} \sum_{i : b \in \mathcal{I}_i} (x_i[b] - \mu[b])^2$$

$$\Rightarrow \hat{C}[b] = \frac{1}{\#(i : b \in \mathcal{I}_i)} \sum_{i : b \in \mathcal{I}_i} (x_i[b] - \hat{\mu}[b])^2 \ . \tag{15}$$

These estimates can be seen as the sample mean and sample variance of the available RSSI measurements for each beacon, without any imputation of missing values. Note that for general Gaussian densities with non-diagonal covariance matrices, the ML estimation with missing values does not have a closed-form solution and requires iterative optimization methods like Newton–Raphson or EM [17], [18].

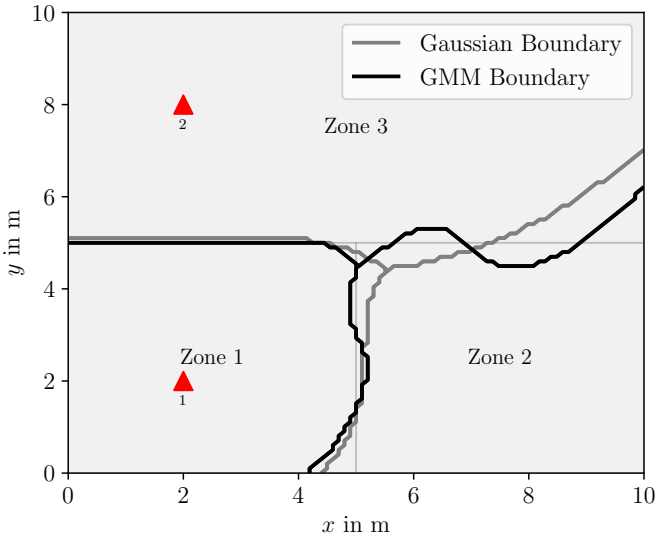


Fig. 2: Toy environment (overall grey area) with three ($N_Z = 3$) zones $\mathcal{Z} = \{Z_1, Z_2, Z_3\}$ (separated by grey straight lines) and two beacon positions (red triangles). Estimated zone boundaries for two classification methods are shown (black and grey irregular curves). Corresponding RSSI space in Fig. 4 with index corresponding to number below triangles.

C. Gaussian Mixture (GM) Estimation

To go beyond naive Bayes and capture more complex distributions of the RSSI data, we extend the model to a GM with K axis-aligned components. Analogous to the single Gaussian case, marginalization over missing dimensions applies to each Gaussian component individually, since sum and integral can be exchanged. The resulting density on the observed dimensions $\mathcal{I}$ is

$$f_{\mathcal{I}}\left(\underline{x} \mid \underline{w}, \{\underline{\mu}_k, \underline{C}_k\}_{k=1}^K\right) = \sum_{k=1}^K w_k \prod_{b \in \mathcal{I}} \mathcal{N}(x[b] \mid \mu_k[b], C_k[b]) \ , \tag{16}$$

where w_k is the weight of the k-th component, with $\sum_{k=1}^K w_k = 1$, $\underline{\mu}_k$ is its mean vector, and $\underline{C}_k$ is its variance vector, see (1).

We estimate the parameters via a modified EM algorithm that handles missing data analogously to the closed-form ML derivation in Section III-B. This can also be derived as a special case of the EM method from [13], as noted in [15]. For the case of weighted measurements, e.g., when the association to the zones is not unique as they have been recorded on the boundary between zones, weighted GM estimation [19] can also easily be introduced.

a) E-Step: For each data point $(\underline{x}_i, \mathcal{I}_i)$, the responsibility γ_{ik} of each component k to each sample i is computed using only the observed dimensions

$$\gamma_{ik} = \frac{w_k \prod_{b \in \mathcal{I}_i} \mathcal{N}(x_i[b] \mid \mu_k[b], C_k[b])}{\sum_{j=1}^K w_j \prod_{b \in \mathcal{I}_i} \mathcal{N}(x_i[b] \mid \mu_j[b], C_j[b])} \ . \tag{17}$$

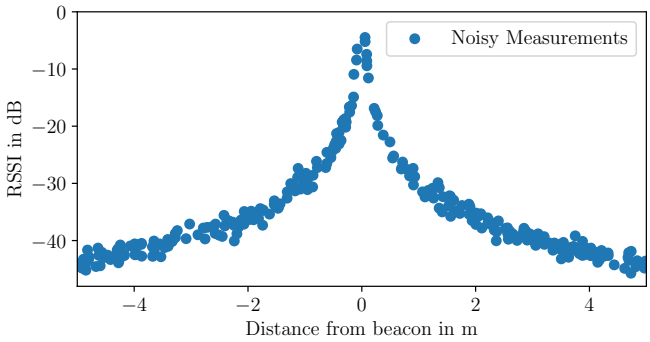


Fig. 3: Toy sensor model: RSSI signal decay with distance. Additive noise.

b) M-Step: Knowledge of the responsibilities or "soft associations" γ_{ik} turns the overall ML problem of all GM parameters into K separate ML problems for the K parameter sets $(w_k, \underline{\mu}_k, \underline{C}_k)$, respectively. This yields the closed-form update rules

$$w_k^+ = \frac{1}{L} \sum_{i=1}^L \gamma_{ik} \ , \tag{18}$$

$$\mu_k^+[b] = \frac{\sum_{i:\, b \in \mathcal{I}_i} \gamma_{ik} \cdot x_i[b]}{\sum_{i:\, b \in \mathcal{I}_i} \gamma_{ik}} \ , \tag{19}$$

$$C_k^+[b] = \frac{\sum_{i:\, b \in \mathcal{I}_i} \gamma_{ik} \cdot (x_i[b] - \mu_k^+[b])^2}{\sum_{i:\, b \in \mathcal{I}_i} \gamma_{ik}} \ . \tag{20}$$

Note that the mean and variance updates for each beacon b use only those samples where beacon b was observed, weighted by the responsibilities, and no imputation is required. E-step and M-step are repeated until the relative change in log-likelihood falls below a threshold ϵ. The parameters are initialized via k-means clustering: cluster centers serve as initial means $\underline{\mu}_k$, cluster variances as initial $\underline{C}_k$, and all weights are set to $w_k = 1/K$.

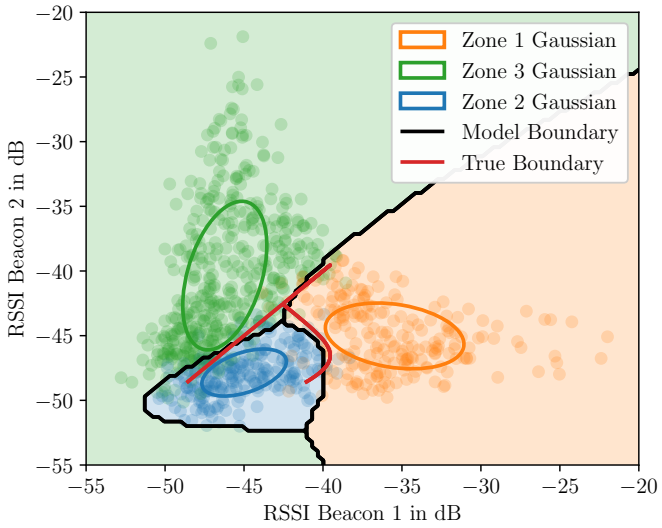
D. Zone Estimation

So far we have estimated the likelihoods $f(\underline{x} \mid Z_z)$ for each zone Z_z. We can use these likelihoods for zone classification by applying Bayes' theorem, as is also done in [16]. First, we need to specify the prior probabilities for each zone, denoted as $f^{\mathrm{p}}(Z_z)$. These priors can be defined as proportional to the zone size, estimated from the training dataset via the number of points per zone, or using some other source of experiential knowledge about the general probability of being in the individual zones. Then, we can compute the posterior probability for each zone given the observed RSSI fingerprint $\underline{x}$ with observed index set $\mathcal{I}$ via

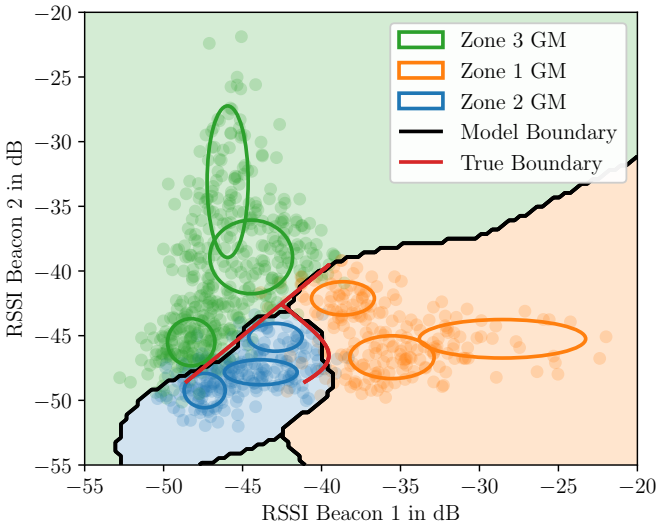
$$f_{\mathcal{I}}(Z_z \mid \underline{x}) = \frac{f_{\mathcal{I}}(\underline{x} \mid Z_z) f^{\mathrm{p}}(Z_z)}{\sum_{z'=1}^{N_Z} f_{\mathcal{I}}(\underline{x} \mid Z_{z'}) f^{\mathrm{p}}(Z_{z'})} \ . \tag{21}$$

If a unique estimate is desired, the maximum a posteriori (MAP) estimate can be used

$$\hat{Z}_z = \arg\max_{Z_z} f_{\mathcal{I}}(Z_z \mid \underline{x}) \ . \tag{22}$$



(a) The single fully correlated Gaussian (ellipse) per zone cannot capture the nonlinear shape of the true density.



(b) Proposed axis-aligned GM ($K = 3$ ellipses). The multiple components approximate the nonlinear density better and yield improved boundaries.

Fig. 4: Toy example decision boundaries (irregular black curves) between zones (orange, green, blue background) in a 2D RSSI space ($N_B = 2$ beacons) using two density models (colored ellipses) learned from data (colored dots): (a) full-covariance Gaussian and (b) axis-aligned mixture model. True decision boundaries indicated by red curves.

E. Why Axis-Aligned Gaussian Components?

Computation is conceptually and computationally much easier with the proposed axis-aligned Gaussians during the training phase. The main reason for axis-aligned GM components, however, is much better efficiency during evaluation (zone estimation). Assume we had fully populated $\mathbb{R}^{N_B \times N_B}$ covariance matrices. Then, for every different pattern of missing values, i.e., for every different index set $\mathcal{I}$, which could differ for each individual measurement, we would have to compute the inverse of the covariance matrix with the missing dimensions marginalized out to evaluate the Gaussian density, which is computationally expensive and not feasible for all real-time applications. Of course, the individual Gaussian components have decreased flexibility if they are restricted to diagonal covariances, but that can easily be compensated with a few more overall components K, as we will demonstrate in Section IV-B1. In general, this does not appear to affect the adaptability and expressivity of the model. Mixture models of axis-aligned components have been applied in various contexts very successfully, for example, in Gaussian process regression [20], kernel methods with axis-aligned basis functions [21], and tensor decomposition-based methods [22], [23], [24].

Note that our method is *not* a naive Bayes classifier. While our proposed GM has axis-aligned and therefore independent or naive components, their *sum* can very well capture arbitrary dependencies in the N_B-dimensional joint RSSI space, see Fig. 1.

IV. Evaluation

We evaluate the proposed method in two scenarios: a synthetic toy example that allows visual inspection of the estimated densities[1] and a real-world warehouse dataset with 21 beacons and missing data.

A. Toy Example: Axis-Aligned GM vs. Correlated Gaussian

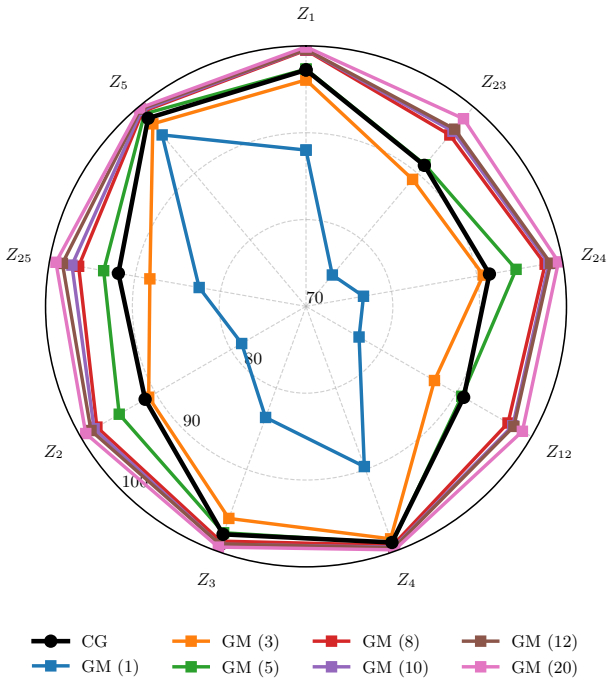
We construct a toy environment consisting of a square domain of $10\,\mathrm{m} \times 10\,\mathrm{m}$ partitioned into $N_Z = 3$ zones and with $N_B = 2$ beacons, as shown in Fig. 2. Each beacon's RSSI characteristic is modeled by a logarithmic decay with additive Gaussian noise, as illustrated in Fig. 3.

Using only two beacons makes the classification task challenging but allows complete visualization in the two-dimensional RSSI space of possible fingerprints. In that space, the true zone densities are nonlinear due to the logarithmic sensor model, and a single correlated Gaussian per zone cannot capture this shape accurately. Fig. 4a shows the MAP (22) decision boundaries obtained with a fully correlated multivariate Gaussian model. The elliptical density contours are a poor fit for the curved true densities, leading to suboptimal boundaries.

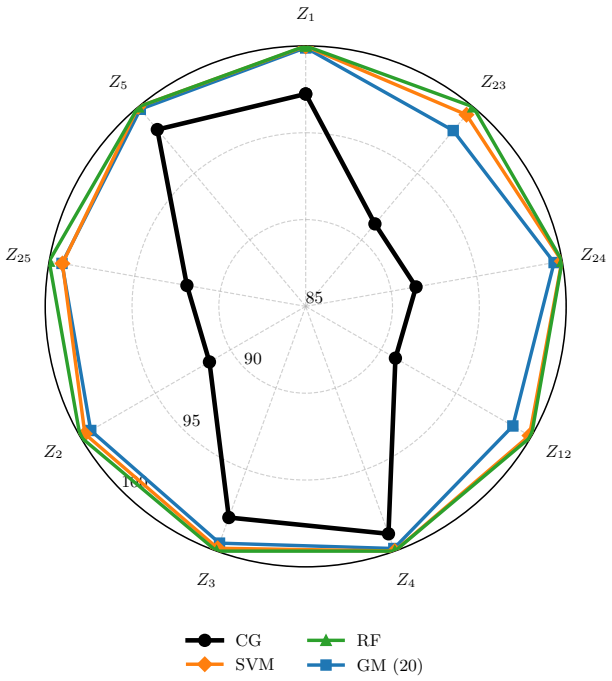
In contrast, the proposed axis-aligned GM adapts to the nonlinear density shape by placing multiple components along the curved region. Fig. 4b shows the resulting MAP decision boundaries using a GM model with $K = 3$ axis-aligned components for every zone. Despite the individual components being uncorrelated and independent, the mixture captures the nonlinear curved shape of the density and produces tighter, more accurate decision boundaries.

Table I reports the F_1 scores from 10 test runs for both methods across different numbers of beacons. Even with only two beacons, the axis-aligned mixture ($89.1 \pm 1.2\,\%$) outperforms

[1]Code is available on GitHub (https://github.com/KIT-ISAS/mixture_positioning_FUSION26) and CodeOcean (linked on the IEEE Xplore page).



(a) Comparison of a fully correlated single Gaussian (CG) and the proposed axis-aligned GM with varying numbers of components K.



(b) Comparison of the proposed fast axis-aligned GM with $K = 20$ components to the slower Random Forest (RF) and support vector machine (SVM) classifiers that rely on imputation for missing data.

Fig. 5: Per-zone F_1 score (radius) for the different zones (angles) on the real-world dataset, obtained from various classification methods (colored lines). Classification zones correspond to Fig. 6.

the fully correlated Gaussian ($88.5 \pm 0.9\,\%$). As more beacons are added, overall accuracy increases and the advantage of the mixture model persists, reaching ($95.6 \pm 0.9\,\%$) with six beacons. Notably, the mixture model achieves this without using any correlation entries in the individual covariance matrices and without requiring matrix inversions or solving linear equation systems to evaluate the Gaussian for different missing data patterns.

TABLE I: F_1 score (in percent) on the toy example for a fully correlated Gaussian (CG) vs. the proposed axis-aligned GM with $K = 3$. First row corresponds to Fig. 4.

Beacon count	CG	GM (ours)
2	88.5 ± 0.9	89.1 ± 1.2
4	95.1 ± 1.2	95.5 ± 0.9
6	94.2 ± 1.2	95.6 ± 0.9

B. Real-World Dataset

We evaluate the proposed method on real RSSI data collected in a warehouse environment with $N_B = 21$ Bluetooth beacons deployed across multiple zones. The dataset exhibits missing values (in approximately $0.6\,\%$ of fingerprints), as not all beacons are received in every time window. The dataset consists of over 210,000 fingerprints, therefore the results show the performance of the methods with enough data. The physical setup is shown in Fig. 6. We perform an 80/20 train/test split.

1) Axis-Aligned GM vs. Correlated Gaussian: We first compare the proposed axis-aligned GM to a fully correlated multivariate Gaussian baseline that handles missing data via submatrix inversion. Fig. 5a shows the per-zone F_1 score for different numbers of mixture components. With a sufficient number of components, the axis-aligned GM matches ($K = 3 \ldots 5$) and exceeds ($K \geq 8$) the correlated Gaussian baseline (black) across most zones, again without requiring any matrix inversions at prediction time.

2) Comparison to Alternative Classifiers: We additionally compare to Random Forest (RF) (with a maximum depth of 50 using 100 trees) and a Support Vector Machine (SVM) classifier (with RBF kernel and one-vs-rest (OvR) decision function), both of which require imputation of missing values prior to classification that is handled by an Iterative Imputer implemented in [25]. Although the RF and SVM methods are not highly optimized, they serve as representatives for their class due to their already high performance on the example. Fig. 5b shows the results. The proposed axis-aligned GM reaches a comparable level of performance to RF and SVM, despite requiring no imputation and using diagonal per-component covariance matrices. This demonstrates that the proposed method can compete with highly performant classifiers while being computationally simpler and statistically principled in its handling of missing data.

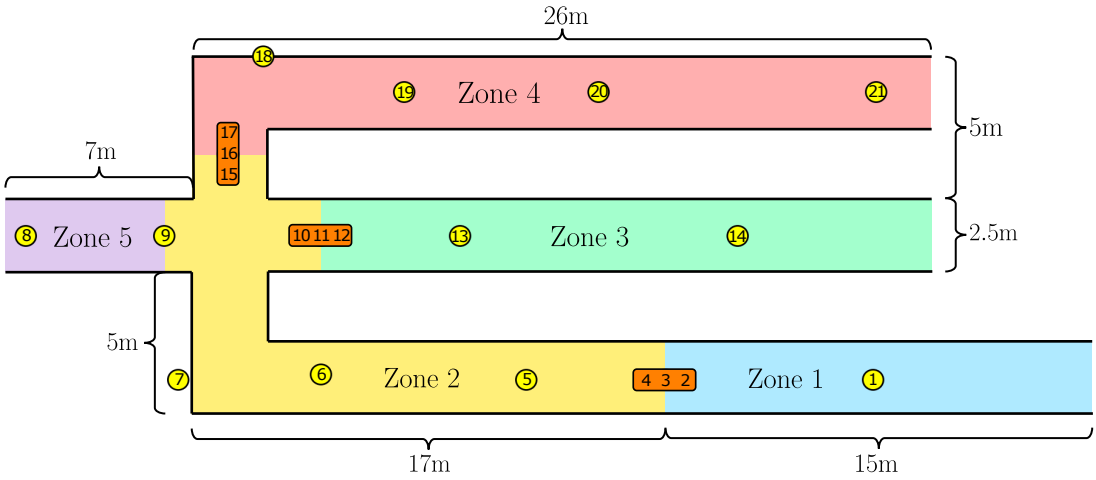


Fig. 6: Layout of the real-world warehouse environment (as it is used in the mentioned ZIM-Project) with beacon positions (yellow circles) and zone boundaries. The orange blocks symbolize the transition zones that are also predicted in the evaluation.

C. Modified Real-World Dataset

To investigate the performance with i) less data, we use a training fraction of only $5\,\%$, and ii) more randomly missing data, we delete one measurement from some fingerprints. Results can be seen in Fig. 7, where a $100\,\%$ missing fraction means every fingerprint lacks one out of 21 RSSI values, thus a total of about $4.8\,\%$ of values are missing. The results indicate that the models based on marginalization keep their performance in missing-data situations just like imputation-based methods, while not needing a trained imputer to reproduce missing values.

V. Conclusion

We presented a method for zone-based indoor localization using axis-aligned GM models that handles incomplete RSSI data in a principled way. The key insight is that axis-aligned components reduce marginalization over missing dimensions to simply dropping factors from a product, eliminating the need for imputation and matrix inversions entirely. This yields a per-prediction cost of $\mathcal{O}(N_B)$ instead of the $\mathcal{O}(N_B^3)$ for full-covariance models, making the method well suited for resource-constrained embedded devices. We derived closed-form ML parameter estimates and a corresponding EM algorithm that naturally incorporates partially observed training samples without requiring surrogate values.

We demonstrated the capability of the method and the effectiveness of using more components to compensate for missing correlations. On real-world data, we additionally demonstrated that the proposed method can keep up with state-of-the-art machine learning methods.

A limitation of the approach is that an appropriate number of mixture components is needed to approximate complex density shapes, and selecting this number currently requires cross-validation. Furthermore, while marginalization is statistically principled, it can lead to conservative predictions when many dimensions are missing simultaneously, as the remaining observed dimensions may carry limited discriminative information.

In the future, we plan to incorporate uncertainty estimates of the individual RSSI measurements, include a dynamical model with state transition (zone transition) probabilities, make detailed runtime comparisons, and implement the prediction on an embedded system.

Acknowledgements

The authors thank Felix Rohn for providing the real-world datasets, the evaluation pipeline, and reference implementations.

References

[1] Ramsey Faragher and Robert Harle. “Location fingerprinting with bluetooth low energy beacons”. In: *IEEE journal on Selected Areas in Communications* 33.11 (2015), pp. 2418–2428.

[2] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[3] Olivier Delalleau, Aaron Courville, and Yoshua Bengio. “Efficient EM training of Gaussian mixtures with missing data”. In: *arXiv preprint arXiv:1209.0521* (2012).

[4] Batyrbek Zholamanov, Ahmet Saymbetov, Madiyar Nurgaliyev, Askhat Bolatbek, Gulbakhar Dosymbetova, Nurzhigit Kuttybay, Sayat Orynbassar, Ainur Kapparova, Nursultan Koshkarbay, and Ömer Faruk Beyca. “RSSI Fingerprint-Based Indoor Localization Solutions Using Machine Learning Algorithms: A Comprehensive Review”. In: *Smart Cities* 8.5 (2025), p. 153.

[5] Navneet Singh, Sangho Choe, and Rajiv Punmiya. “Machine Learning Based Indoor Localization Using Wi-Fi RSSI Fingerprints: An Overview”. In: *IEEE Access* 9 (2021), pp. 127150–127174. DOI: 10.1109/ACCESS.2021.3111083.

[6] Zeynep Turgut, Serpil Üstebay, Gülsüm Zeynep Gürkaş Aydın, and Ahmet Sertbaş. “Deep learning in indoor localization using WiFi”. In: *International Telecommunications Conference: Proceedings of the ITelCon 2017, Istanbul*. Springer. 2018, pp. 101–110.

[7] Marwan Alfakih, Mokhtar Keche, Hadjira Benoudnine, and Abdelkrim Meche. “Improved Gaussian mixture modeling for accurate Wi-Fi based indoor localization systems”. In: *Physical Communication* 43 (2020), p. 101218. DOI: https://doi.org/10.1016/j.phycom.2020.101218.

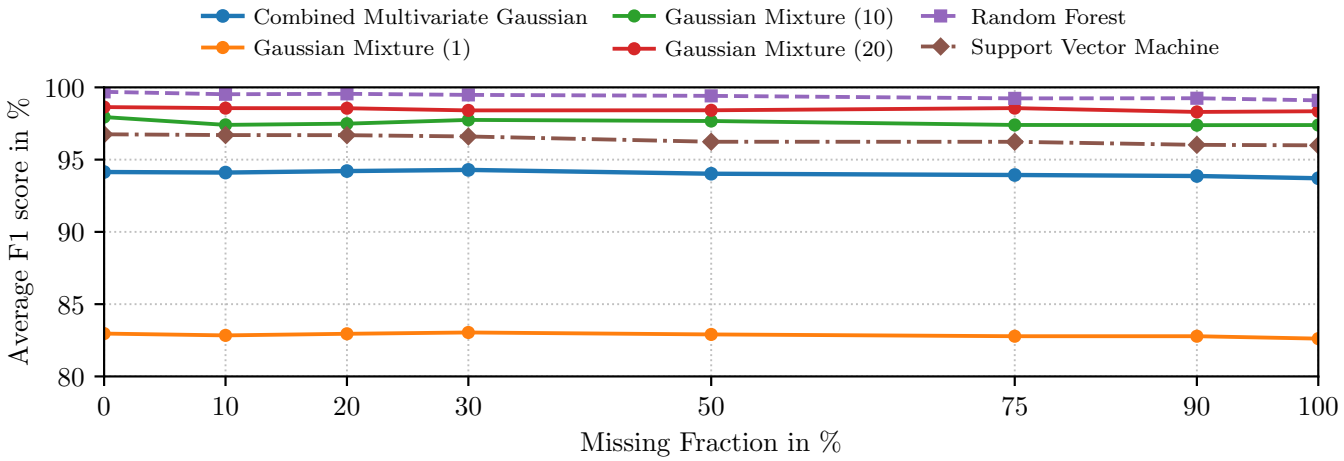


Fig. 7: Averaged F_1 score over all predicted zones. Missing fraction describes the fraction of fingerprints in which one random RSSI value was removed.

[8] Yan Li, Simon Williams, Bill Moran, Allison Kealy, and Guenther Retscher. "High-Dimensional Probabilistic Fingerprinting in Wireless Sensor Networks Based on a Multivariate Gaussian Mixture Model". In: *Sensors* 18.8 (2018). DOI: 10.3390/s18082602.

[9] Parvin Malekzadeh, Mohammad Salimibeni, Mohammadamin Atashi, Mihai Barbulescu, Konstantinos N. Plataniotis, and Arash Mohammadi. "Gaussian Mixture-based Indoor Localization via Bluetooth Low Energy Sensors". In: *2019 IEEE SENSORS*. 2019, pp. 1–4. DOI: 10.1109/SENSORS43011.2019.8956950.

[10] Minhui Luo, Jin Zheng, Wei Sun, and Xing Zhang. "WiFi-based Indoor Localization Using Clustering and Fusion Fingerprint". In: *2021 40th Chinese Control Conference (CCC)*. 2021, pp. 3480–3485. DOI: 10.23919/CCC52363.2021.9549410.

[11] Beenish A. Akram, Ali H. Akbar, and Omair Shafiq. "HybLoc: Hybrid Indoor Wi-Fi Localization Using Soft Clustering-Based Random Decision Forest Ensembles". In: *IEEE Access* 6 (2018), pp. 38251–38272. DOI: 10.1109/ACCESS.2018.2852658.

[12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[13] Zoubin Ghahramani and Michael Jordan. "Supervised learning from incomplete data via an EM approach". In: *Advances in neural information processing systems* 6 (1993).

[14] Yan Li, Simon Williams, Bill Moran, Allison Kealy, and Guenther Retscher. "High-dimensional probabilistic fingerprinting in wireless sensor networks based on a multivariate Gaussian mixture model". In: *Sensors* 18.8 (2018), p. 2602.

[15] John Canny. "Collaborative filtering with privacy via factor analysis". In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002, pp. 238–245.

[16] Michal Stepánek, Jirí Franc, and Václav Kůs. "Modification of Gaussian mixture models for data classification in high energy physics". In: *Journal of Physics: Conference Series*. Vol. 574. 1. 2015, p. 012150.

[17] Irene Monahan Trawinski and R. E. Bargmann. "Maximum Likelihood Estimation with Incomplete Multivariate Data". In: *The Annals of Mathematical Statistics* 35.2 (1964), pp. 647–657. URL: http://www.jstor.org/stable/2238516 (visited on 02/06/2026).

[18] E. M. L. Beale and R. J. A. Little. "Missing Values in Multivariate Analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 37.1 (Dec. 2018), pp. 129–145. DOI: 10.1111/j.2517-6161.1975.tb01037.x.

[19] Daniel Frisch and Uwe D. Hanebeck. "Gaussian Mixture Estimation from Weighted Samples". In: *Proceedings of the 2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2021)*. Karlsruhe, Germany, Sept. 2021. DOI: 10.1109/MFI52462.2021.9591161.

[20] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Nov. 2005. DOI: 10.7551/mitpress/3206.001.0001.

[21] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning". In: *The Annals of Statistics* 36.3 (2008), pp. 1171–1220. DOI: 10.1214/009053607000000677.

[22] Felix Govaers, Bruno Demissie, Altamash Khan, Martin Ulmke, and Wolfgang Koch. "Tensor Decomposition-Based Multitarget Tracking in Cluttered Environments". In: *Journal of Advances in Information Fusion* 14.1 (2019), p. 86. URL: https://isif.org/media/tensor-decomposition-based-multitarget-tracking-cluttered-environments.

[23] Daniel Frisch and Uwe D. Hanebeck. "Fokker-Planck Prediction on the Cylindric Manifold using Tensor Decomposition of a Regular Grid". In: *17th Symposium Sensor Data Fusion: Trends, Solutions, Applications (SDF 2025)*. Bonn, Germany, Nov. 2025. DOI: 10.1109/SDF67080.2025.11330240.

[24] Daniel Frisch and Uwe D. Hanebeck. "Interpolation of Probability Densities on a Grid in Tensor Decomposition Representation". In: *Proceedings of the 29th International Conference on Information Fusion (FUSION 2026)*. Trondheim, Norway, June 2026.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.