Mokise and Myburgh
EURASIP Journal on Advances in Signal Processing (2023) 2023:99
https://doi.org/10.1186/s13634-023-01059-2

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Model-based clustering of multipath propagation in powerline communication channels

Kealeboga L. Mokise[1*] and Herman C. Myburgh[1]

*Correspondence:
kealeboga.mokise@up.ac.za

[1] Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, Gauteng, South Africa

Abstract

Powerline communication (PLC) channels are known to exhibit multipath propagation behaviour. The authors present a model-based framework to address the challenge of clustering multipath propagation components (MPCs) in PLC channels for indoor low-voltage (LV) environments. The framework employs a range of finite-mixture models (FMMs), including the gamma mixture model, the inverse gamma mixture model, the Gaussian mixture model, the inverse Gaussian mixture model, the Nakagami mixture model, the inverse Nakagami mixture model (INMM) and the Rayleigh mixture model, to identify clusters of MPCs. A measurement campaign of an unknown indoor LV PLC channel is conducted to obtain a channel response. From the channel response, the delay and magnitude parameters of the MPCs are extracted using the space-alternating generalised expectation maximisation algorithm adopted only for these parameters. A maximum likelihood approach and the expectation–maximisation algorithm are employed to fit the FMMs to the MPC delay-magnitude dataset to cluster MPCs in the delay domain. The results of the model-fitting process are then evaluated using the corrected Akaike information criterion (AICc), which enables a fair comparison of the candidate models over the feasible and finite range of clusters. A novel algorithm is introduced for estimating the feasible and finite range of clusters using the extracted delay and magnitude MPC parameters. The AICc's ranking results show that the INMM model provides the best fit. Davies–Bouldin (DB) and Calinski–Harabasz (CH) indexes are used to compare the model-based clustering approach to the conventional distance-based clustering methods. Validation results show that CH and DB indexes closely agree in the optimal number of MPC clusters for model-based clustering, which corresponds to the most within-cluster compactness of MPCs and to the most between-cluster separation in the delay domain.

Keywords: Finite-mixture model, Powerline communication channels, Akaike information criterion, Model-based clustering

1 Introduction

The rapid advancement of communication technologies has resulted in an increased demand for spectral resources. With wider bandwidths and higher transceiver passband frequencies, radio signals are increasingly subjected to scattering in the propagation

environment. This scattering results in multipath propagation and the formation of clusters of multipath propagation components (MPCs) [1]. Given the increased usage of spectral resources and increased scattering, it is crucial that channel models accurately account for the clustering of scatterers. Some examples of such models include the COST259 direction channel model (DCM) [2], designed for third and fourth-generation systems, the Saleh–Valenzuela (SV) channel model [3], which models the arrival of MPCs in clusters for indoor wideband (WB) wireless transmission systems, and the geometry-based stochastic model (GBSM) [4], which is a cluster-based channel model adopted for fifth-generation systems.

In channel modelling, a cluster refers to a group of MPCs which share similar parameters such as delay (τ), azimuth angle of arrival (AOA), azimuth angle of departure (AOD), elevation angle of departure (EOD) and elevation angle of arrival (EOA) [5, 6]. However, there is no universal definition of a cluster, which means the cluster definition and clustering results depend on the clustering method used. In earlier works of MPC clustering, clusters were manually identified from extracted multipath parameters, such as delay and magnitude, by using visual observation methods [7–9]. Manual clustering performs well if inter-cluster and intra-cluster parameters are clearly distinct; however, when clusters overlap, manual methods result in erroneous cluster identification. Conventionally, automatic clustering is performed using distance-based methods such as k-means, kPowerMeans (kPM) [10] and fuzzy-c means [11]. k-means identifies clusters by finding the distances between MPCs, and kPM offers an improvement of clustering results by considering the power of the multipath components. Fuzzy-c means is a soft-decision alternative to k-means where a fuzzifier parameter is considered when computing the distances between the MPCs. Fuzzy-c means typically outperforms both k-means and kPM when cluster centroids are randomly initialised. However, the deterministic initialisation of centroids leads to similar clustering results as the kPM with a slight improvement over k-means. Once a cluster is identified, a cluster validity index (CVI) is used to select the best partition of clusters for the dataset. CVIs such as those of Dunn [12], Davies–Bouldin [13] and Calinski–Harabasz [14] select the best partition by computing the separation and compactness of clusters. There is no single universally superior CVI, rather, different CVIs can select different best cluster partitions for a particular dataset. Arbelaitz et al. [15] conducted an extensive study of 30 cluster validity indexes to investigate cluster partition selection through the separation and compactness of clusters. Their study highlights that optimal cluster validation does not exist. The selection of best the cluster partition depends on the dataset configuration. Ultimately, CVIs offer insight into the effectiveness of the clustering method over the considered range cluster partitions.[1]

Model-based clustering assumes that the dataset distribution can be described by a multimodal probability density function (PDF). This is a linear combination of K finite unimodal PDFs, which are component densities. Assuming a mixture model with a finite

[1] Ultimately cluster validation would give insight as to which clustering method gives favourable and reliable clustering results, for example, over the considered range of cluster partitions the max of the Calinski–Harabasz index value indicate the best cluster partition. When comparing two clustering methods using the Calinski–Harabasz index, if one method consistently results in high index values over the considered of cluster partitions, that would mean such method would give more favourable and reliable best cluster partition.

number of component densities, the clustering problem becomes one of estimating the parameters of the component densities, whereby each component density represents a cluster, and the posterior probabilities of each MPC determine its cluster membership to one of the K component densities [16]. The authors conducted a measurement campaign of an indoor low-voltage (LV) powerline network, and used the space-alternating generalised expectation–maximisation (SAGE) [17] algorithm to extract the delay-magnitude MPC parameters. From the extracted parameters, the expectation–maximisation (EM) method was used to estimate the parameters of the component densities.

Powerline communication (PLC) channels are known to exhibit multipath propagation, specifically in indoor LV PLC channels, in which a propagating signal experiences multipath propagation due to discontinuities and an impedance mismatch of loads [18–20]. The authors considered an indoor LV power network in which electrical outlets and loads were not randomly distributed, but concentrated at regular intervals. This, is typical of most indoor LV networks in residential and commercial buildings. Therefore, a signal propagating in such a medium would experience multiple reflections due to the concentrated group of discontinuities and loads. This, would result in a cluster of MPCs.[2] A signal propagation path in PLC channels typically requires a direct connection between discontinuities and loads, such that a direct signal path exists between the transmitter and the receiver.[3] A cluster resulting from such a channel would have a dominant component and additional components with time-decaying magnitudes. It is therefore expected to exhibit a positively skewed distribution of MPCs.

The first use of model-based multipath clustering appears in [21, 22], where a Gaussian mixture model (GMM) was used to identify clusters in a wireless propagation channel. The authors postulate that Gaussian finite-mixture models (FMMs) are well-suited for MPC clustering since the scattering property of the wireless channel obeys a Gaussian distribution. However, this assumes that the signal only experiences diffuse scattering in the channel. In real-world channels, it is often reported in the literature that a cluster typically has a dominant component and additional MPCs with time-decaying magnitudes. This is also an inherent property of cluster-based models such as the SV model. Therefore, it can be assumed that, in general, the power delay profile (PDP) of the channel will exhibit a positively skewed distribution. The clusters within the PDP will also exhibit a positively skewed distribution. A positively skewed distribution is best described by long-tail models such as gamma, Nakagami and log-normal. To the best of the authors' knowledge, the work in this paper presents the first investigation of MPC cluster identification in PLC channels. Moreover, this work introduces the first application of the gamma mixture model (G_{γ}MM), the inverse gamma mixture model (IG_{γ}MM), the Nakagami mixture model (NMM), the inverse Nakagami mixture model (INMM) and the inverse Gaussian mixture model (IGMM) to MPC clustering in PLC applications. The candidate models adopted are well-suited for describing positively skewed distributions.

[2] This is synonymous with a propagating wireless signal that encounters a group of scatterers on its path, resulting in a cluster of multipath components.

[3] This is synonymous with a line-of-sight path in a wireless channel.

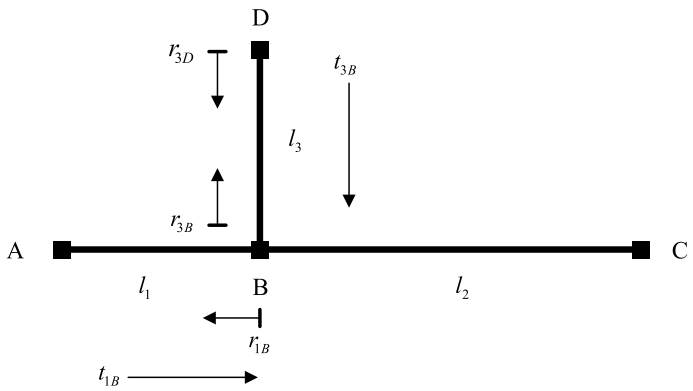

Fig. 1 Multipath behaviour in a simple T-network PLC channel

The rest of the paper is organised as follows: Section 2 presents the problem formulation for MPC clustering for an indoor LV PLC channel with branch-connection clusters, and discusses considerations and assumptions imposed on such a channel. Section 3 discusses the distance-based and model-based clustering solutions. With the exception of the GMM, an in-depth description of the EM procedure for each FMM clustering solution is provided. The model selection and cluster validation procedures are described in Sect. 3. Section 4 presents the results of the channel measurements, the estimation of delay and magnitude parameters of the MPCs using the SAGE algorithm, model selection and cluster validation procedures. The paper is concluded in Sect. 5.

2 Multipath cluster problem in powerline channels

Several models in the literature describe the channel transfer characteristics of the PLC channel. These models are commonly categorised as parametric [18, 23] and deterministic [24, 25]. Parametric models are derived using a data-fitting approach to estimate the model parameters, while deterministic models are derived from transmission-line theory, which requires detailed information about the physical medium. Both modelling approaches consider the multipath propagation behaviour of the PLC channel. Figure 1 shows a simple T-network PLC channel, which consists of a single branch between point D and point B, and a direct path from A to B to C. The line segments have lengths l_1, l_2 and l_3 with characteristic impedances Z_{l1}, Z_{l2} and Z_{l3}, respectively. When a signal propagates in the network, it does not only propagate along the direct path from A to B to C. Reflections of the transmitted signal are also present. Terminals A and C are assumed to be matched. Therefore, the only points for reflections to occur are B and D, with the reflection factors denoted by r_{1B}, r_{3B} and r_{3D}, and the transmission factors denoted by t_{1B} and t_{3B}.

The signal's direct path is given as $A \rightarrow B \rightarrow C$. The second path is given as $A \rightarrow B \rightarrow D \rightarrow B \rightarrow C$. The Nth path is given as $A \rightarrow B(\rightarrow D \rightarrow B)^{N-1} \rightarrow C$.

Figure 2 shows an indoor LV network where (■) is a branch, (◯) is an open connection, and (•) is a load connected in the network. When a signal propagates from A to E, it will experience transmission and reflection at each branch and connection. The r_{2B} signal component will propagate into branch-connection cluster 1 (BC_{C1}), in which it will experience multipath propagation similar to Fig. 1 at each branch and connection. At some point, a

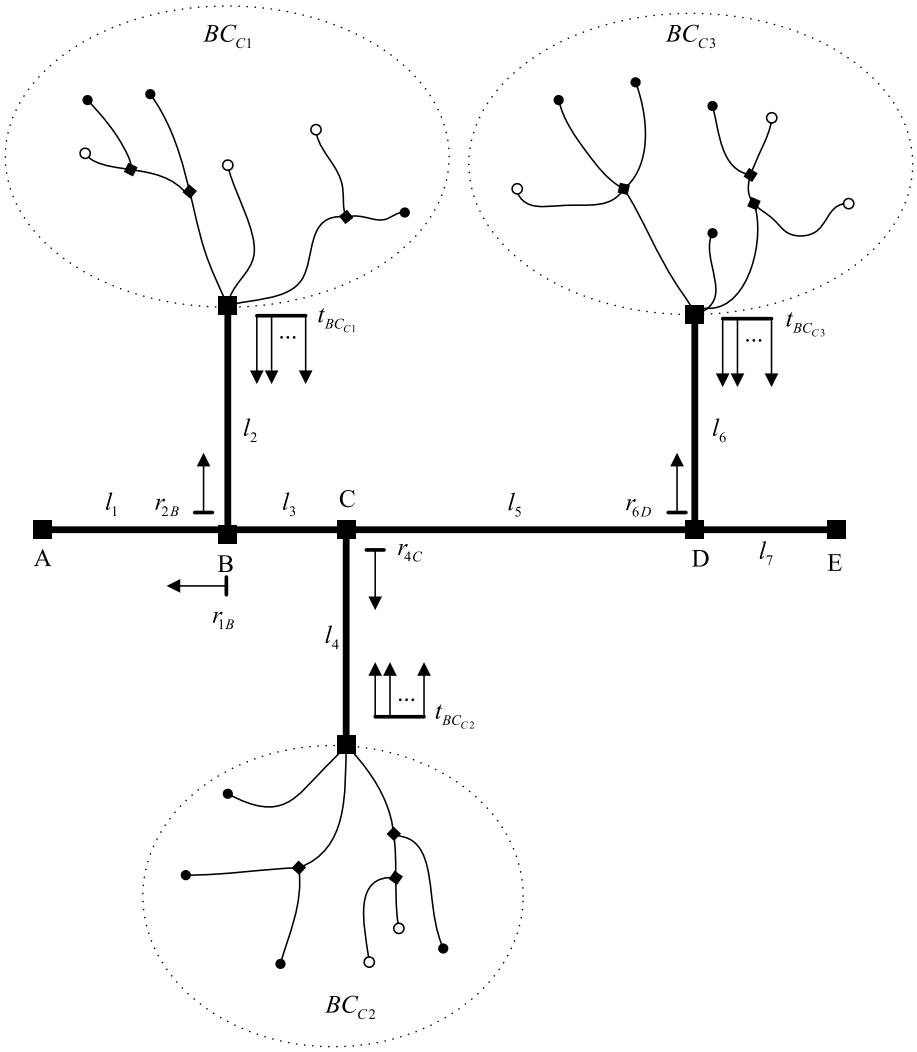

Fig. 2 LV network with branch-connection clusters

finite number of signal components with time-decaying magnitudes will be reflected back from BC_{C1}, as shown by $t_{BC_{C1}}$ in the figure, which is a cluster of MPCs. The same principle would apply for BC_{C2} and BC_{C3} which results in MPC clusters $t_{BC_{C2}}$ and $t_{BC_{C3}}$, respectively. For an input sounding signal $g(t)$, the measured PLC channel response $y(t)$ is given by the convolution

$$y(t) = g(t) * h(t) + n(t), \tag{1}$$

where $h(t)$ denotes the PLC channel impulse response given as

$$h(t) = \sum_{l=1}^{L} \sum_{m=1}^{M_l} \alpha_{m,l} e^{-j2\pi f_c \tau_{m,l}} \delta\left(t - \tau_{m,l}\right), \tag{2}$$

where $l = 1, 2, \ldots, L$ denotes the cluster number, $m = 1, 2, \ldots, M_l$ denotes the MPC in the lth cluster, and $\alpha_{m,l}$ and $\tau_{m,l}$ denote the magnitude and delay of the mth MPC in the

lth cluster. Connected loads may change over long periods of time, which means the channel transfer characteristics can change over the same time periods. However, even in that case, the branching would remain the same, which means the MPC clustering behaviour would still hold. Therefore $h(t)$ in (2) is considered to be a wide-sense stationary process. The sounding signal $g(t)$ is described in more detail in Sect. 4.1.

Background noise and impulsive noise interferences in PLC channels are the result of the corona effect, switching of loads and the electric arc or crosstalk between powerline cables. Background noise can be modelled as additive white Gaussian noise (AWGN), since it is wideband and has low power spectral density (PSD), while impulsive noise occurs in bursts and exhibits high PSD. In general, the noise term $n(t)$ in (1) can be considered a sum of background noise and impulsive noise. However, impulsive noise is typically observed in the tens of megahertz range [26]. In this study, measurements are conducted in the hundreds of megahertz carrier frequency. As such, the effects of impulsive noise can be safely avoided, and $n(t)$ in (1) can be considered as AWGN.

3 Multipath clustering solutions for powerline channels

3.1 Distance-based clustering solution

Multipath component estimation from channel measurements will usually result in L clusters, where each cluster consists of finite M_l MPCs such that $m_l = 1, 2, \ldots, M_l$. Each cluster will have a centre c_i such that $c_i = 1, 2, \ldots, C_L$. The k-means clustering method is a hard-partitioning method that uses a distance metric to minimise the distance sum of the respective m_l over all C_L and assigns m_l to the lth cluster with the minimised c_i. The centroids are then re-estimated by averaging the cluster assigned MPCs. This process iterates until some predefined accuracy stop criterion is reached. The distance metrics commonly used include the squared Euclidean distance (SED), the joint squared Euclidean distance (JSED) and the multipath component distance (MCD). The JSED and MCD allow for clustering MPCs jointly through time and angle parameters, and the SED can cluster MPCs only one dimension at a time. In [5], the authors compare SED, JSED and MCD using the spatial channel model for multiple-input multiple-output (SCM-MIMO) under different angular spreads. The results show that MCD obtain the best performance as the number of incorrectly clustered MPCs decreases only slightly for larger angular spreads. This was mainly attributed to the robust scaling and joint clustering of MPC parameters, which are attributes lacking in the other metrics. The MCD distance metric between the ith and jth MPCs is given as

$$\mathrm{MCD}_{ij} = \sqrt{\left\|\mathrm{MCD}_{Rx,ij}\right\|^2 + \left\|\mathrm{MCD}_{Tx,ij}\right\|^2 + \mathrm{MCD}_{\tau,ij}{}^2}, \tag{3}$$

where

$$\mathrm{MCD}_{Tx/Rx,ij} = \frac{1}{2}\left|\begin{pmatrix}\sin(\theta_i)\sin(\varphi_i)\\ \sin(\theta_i)\cos(\varphi_i)\\ \cos(\varphi_i)\end{pmatrix} - \begin{pmatrix}\sin(\theta_j)\sin(\varphi_j)\\ \sin(\theta_j)\cos(\varphi_j)\\ \cos(\varphi_j)\end{pmatrix}\right|, \tag{4}$$

$$\mathrm{MCD}_{\tau,ij} = \frac{\left|\tau_i - \tau_j\right|}{\Delta\tau_{\max}} \cdot \frac{\tau_{\mathrm{std}}}{\Delta\tau_{\max}}. \tag{5}$$

$\Delta\tau_{\max}$ denotes the maximum difference of the delay spread, τ_{std} denotes the standard deviation of the delay spread, θ_j and φ_j denote the azimuth and elevation, respectively. The MCD, as in (3), is computed as a Euclidean norm vector, which can be interpreted as a hyper-sphere in the normalised distances in the delay and angle domains. MPCs of powerline channels are clustered in the delay domain. Therefore, due to the lack of azimuth and elevation parameters, a distance metric is considered given as

$$\text{MCD}_{ij} = \sqrt{\text{MCD}_{\tau,ij}^2}, \tag{6}$$

which is a reduced version of (3) from a multi-dimensional parameter space to only the delay domain. This is comprehensible since the parameter estimator in the delay domain is more stable and robust than in the angular domain due to its non-periodicity. Therefore, (6) computes the normalised absolute distance between the delays of the corresponding MPCs scaled by the normalised delay spread.

3.2 Model-based clustering solution

Unlike distance-based clustering solutions, model-based clustering allows an MPC to belong to different clusters at the same time, but with different probabilities, where the highest probability indicates the cluster it belongs to. This method provides more flexibility, but at the cost of computation complexity. In model-based clustering using FMMs, the MPCs are described by a vector $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in C^d$ denotes the parameters of the ith MPC, d denotes the dimension of $\mathbf{x}_i$ and N denotes the number of samples. The mixture model is a weighted sum of a finite K component distributions, which is expressed as

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \zeta_k p(x_i|\theta_k). \tag{7}$$

Each x_i is assumed to be generated by one of the K component distribution $p(x_i|\theta_k)$, with a parameter set θ_k, and ζ_k denotes the priors of the component distributions, which satisfies the constraint $\sum_{k=1}^{K} \zeta_k = 1$. For a set of N multipath components constituting dataset $\mathbf{X}$, the goal is to find the set of parameters Θ, which maximises the likelihood of the FMM given $\mathbf{X}$. Assuming the N samples are identically and independently distributed (i.i.d), the FMM likelihood $\mathcal{L}(\Theta|X) = p(X|\Theta)$ is expressed as

$$p(\mathbf{X}|\Theta) = \prod_{i=1}^{M} p(x_i|\Theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} \zeta_k p(x_i|\theta_k). \tag{8}$$

However, direct optimisation of (8) is impossible because the expression is a nonlinear function of the parameter set Θ. Therefore, maximum likelihood (ML) estimates of the parameters are obtained using the EM algorithm, which iterates through parameter estimates and attempts to find the set of parameters that maximises the log of the likelihood function, which is expressed as

$$\log \mathcal{L}(\Theta|\mathbf{X}) = \sum_{i=1}^{M} \log \sum_{k=1}^{K} \zeta_k p(x_i|\theta_k). \tag{9}$$

The EM algorithm obtains the ML estimates by finding the expectation of the complete-dataset log-likelihood with respect to the unobserved dataset $\mathbf{Y}$, given $\mathbf{X}$ and Θ, which is the Q function, expressed as

$$Q\left(\Theta^{t+1}, \Theta^t\right) = E\left[\log P\left(X, Y|\Theta^{t+1}\right)|X, \Theta^t\right]. \tag{10}$$

In order to solve (9), $\mathbf{X}$ is considered to be an incomplete, but observed dataset. Then, assuming $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$ to be the unobserved dataset, the complete-dataset exists as $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ [16]. The $\mathbf{Y}$ dataset informs which component density generated x_i, simply put, assuming $y_i \in \{1, \ldots, K\}$ for each i, then $y_i = k$ if the ith sample was generated by the kth component density. A new likelihood function, $\mathcal{L}(\Theta|\mathbf{Z}) = \mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y})$, is defined as the complete-dataset likelihood. Then, (9) simplifies to the complete-dataset log-likelihood given as

$$\log\left(\mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y})\right) = \sum_{i=1}^{N} \log\left(P\left(x_i|y_i\right)P(y)\right) = \sum_{i=1}^{N} \log\left(\zeta_{y_i} p\left(x_i|\theta_{y_i}\right)\right). \tag{11}$$

The distribution of the unobserved data is obtained from (11) using Bayes rule, which is expressed as

$$p\left(y_i|x_i, \Theta^t\right) = \frac{\zeta_{y_i}^{(t)} p\left(x_i|\theta_{y_i}^{(t)}\right)}{\sum_{j=1}^{K} \zeta_j^{(t)} p\left(x_i|\theta_j^{(t)}\right)}, \tag{12}$$

where the superscript t denotes the currently estimated parameters. The expression in (12) is the probability that x_i was generated by the kth component density. This is the E-step of the EM algorithm, which will be computed in the same way for all the considered FMMs. More specifically, it is computed using (13).

$$w_k^{(i)} = \frac{\zeta_k^{(t)} p\left(x_i|\theta_k^{(t)}\right)}{\sum_{j=1}^{K} \zeta_j^{(t)} p\left(x_i|\theta_k^{(t)}\right)}. \tag{13}$$

Then, the Q function in (10) simplifies to

$$Q\left(\Theta^{t+1}, \Theta^t\right) = \sum_{k=1}^{K}\sum_{i=1}^{M} \log\left(\zeta_k\right) w_k + \sum_{k=1}^{K}\sum_{i=1}^{M} \log\left(p(x_i|\theta_k)\right) w_k, \tag{14}$$

which needs to be maximised to obtain the ML estimates. The right-hand terms containing ζ_k and θ_k can be maximised independently. This is the M-step of the EM algorithm. ζ_k is obtained as

$$\frac{\partial}{\partial \zeta_k}\left[\sum_{k=1}^{K}\sum_{i=1}^{M} \log\left(\zeta_k\right) w_k + \lambda\left(\sum_k \zeta_k - 1\right)\right] = 0 \tag{15}$$

$$\sum_{i=1}^{N} \frac{1}{\zeta_k} w_k + \lambda = 0, \tag{16}$$

taking the sum of both sides over k gives $\lambda = -N$, then

$$\zeta_k = \frac{1}{M} \sum_{i=1}^{M} {w_k}^{(i)}. \tag{17}$$

The M-step of the EM computation of ζ_k follows the same expression for all the FMMs considered. The computation of the parameters θ_k, which is equivalent to maximising the term with θ_k on the right-hand side of (14), will be different for each FMM.

3.2.1 GMM

For the GMM, each x_i is assumed to be generated by one of the K Gaussian densities with mean μ_k and covariance Σ_k, which is expressed as

$$p(x_i|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{(x_i - \mu_k)^T {\Sigma_k}^{-1} (x_i - \mu_k)}{-2} \right). \tag{18}$$

Finding the ML estimates $\hat{\mu}_k$ and $\hat{\Sigma}_k$ is equivalent to taking partial derivative of the term with θ_k on the right-hand side of (14) with respect to μ_k and Σ_k, and setting them to zero. This results in the closed-form expressions (19) and (20) for the mean and covariance, respectively. GMM is a well-known mixture model in the literature, of which the detailed derivation of (19) and (20) can be found in [16, 27] and the references within.

$$\mu_k = \frac{\sum_{i=1}^{M} x_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}, \tag{19}$$

$$\Sigma_k = \frac{\sum_{i=1}^{M} w_k^{(i)} (x_i - \mu_k)^T (x_i - \mu_k)}{\sum_{i=1}^{M} w_k^{(i)}}. \tag{20}$$

The EM algorithm iterates between the E-step, i.e. (13), and the M-step, i.e. (17), (19), and (20), until it converges to some predefined accuracy. It should be noted that MPC clustering for PLC channels is done in the delay domain only. Therefore, the dimensionality $d = 1$ and the full covariance matrix Σ structure is considered for each component density.

3.2.2 IGMM

For the IGMM, each x_i is assumed to be generated by one of the K inverse Gaussian densities with mean μ_k and shape λ_k expressed as

$$p(x_i|\lambda_k, \mu_k) = \left(\frac{\lambda_k}{2\pi {x_i}^2} \right)^{\frac{1}{2}} \exp\left(\frac{-\lambda_k (x_i - \mu_k)^2}{2{\mu_k}^2 x_i} \right), \quad x_i > 0,\ \mu_k > 0,\ \lambda_k > 0. \tag{21}$$

Finding the ML estimates $\hat{\lambda}_k$ and $\hat{\mu}_k$ is equivalent to taking partial derivatives of the term with θ_k on the right-hand side of (14) with respect to λ_k and μ_k, and setting them to zero. Substituting (21) for $\log\left(p(x_i|\theta_k)\right)$ in (14), the expression is given as

$$\sum_{k=1}^{K}\sum_{i=1}^{M}\left[\frac{1}{2}\left(\log\lambda_k - \log 2\pi - 3\log x\right) - \frac{\lambda_k(x_i-\mu_k)^2}{2{\mu_k}^2 x_i}\right]w_k^{(i)}, \tag{22}$$

therefore taking $\frac{\partial}{\partial\lambda}$ and $\frac{\partial}{\partial\mu}$ of (22) and setting them to zero gives (23) and (24), respectively.

$$\sum_{i=1}^{M}\left(\frac{\lambda_k(x_i-\mu_k)}{{\mu_k}^3}\right)w_k^{(i)} = 0, \tag{23}$$

$$\sum_{i=1}^{M}\frac{1}{2}\left(\frac{1}{\lambda_k} - \frac{(x_i-\mu_k)^2}{x_i{\mu_k}^2}\right)w_k^{(i)} = 0. \tag{24}$$

Solving (23) for μ_k gives the closed-form expression

$$\mu_k = \frac{\sum_{i=1}^{M} x_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} = \bar{x}_i. \tag{25}$$

Then, substituting for μ_k, i.e. $\bar{x}_i$, in (24) and solving for λ_k gives the closed-form expression

$$\lambda_k = \frac{\sum_{i=1}^{M}\left(\frac{(\bar{x}_i)^2 x_i}{(x_i-\bar{x}_i)^2}\right)w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}. \tag{26}$$

The EM algorithm iterates between the E-step given by (13) and the M-step given by (17), (25) and (26), until it converges.

3.2.3 RMM

For the RMM, each x_i is assumed to be generated by one of the K Rayleigh densities with mean σ, which is expressed as

$$p(x_i|\sigma_k) = \frac{x_i}{{\sigma_k}^2}\exp\left(-\frac{{x_i}^2}{2{\sigma_k}^2}\right), \quad x_i > 0,\ \sigma_k > 0. \tag{27}$$

Finding the ML estimate $\hat{\sigma}_k$ is equivalent to taking a partial derivative of the term with θ_k on the right-hand side of (14) with respect to σ_k and setting it to zero. Substituting (27) for $\log\left(p(x_i|\theta_k)\right)$ in (14) gives the expression

$$\sum_{k=1}^{K}\sum_{i=1}^{M}\left[\log x_i - \log{\sigma_k}^2 - \frac{{x_i}^2}{2{\sigma_k}^2}\right]w_k^{(i)}, \tag{28}$$

therefore taking $\frac{\partial}{\partial\sigma}$ of (28) and setting it to zero gives (29).

$$\sum_{i=1}^{M}\left(\frac{{x_i}^2}{{\sigma_k}^3}-\frac{2}{\sigma_k}\right)w_k^{(i)}=0. \tag{29}$$

Solving (29) for σ_k gives the closed-form expression

$$\sigma_k=\sqrt{\frac{\sum_{i=1}^{M}{x_i}^2 w_k^{(i)}}{2\sum_{i=1}^{M}w_k^{(i)}}}. \tag{30}$$

The EM algorithm iterates between the E-step given by (13) and the M-step given by (17) and (30) until it converges to some predefined accuracy.

3.2.4 $G_\gamma MM$

For the G_γMM, each x_i is assumed to be generated by one of the K gamma densities with shape α_k and scale β_k , which is expressed as

$$p(x_i|\alpha_k,\beta_k)=\frac{{x_i}^{(\alpha_k-1)}}{{\beta_k}^{\alpha_k}\Gamma(\alpha_k)}\exp\left(-\frac{x_i}{\beta_k}\right),\quad x_i>0,\ \alpha_k>0,\ \beta_k>0. \tag{31}$$

Obtaining the ML estimates $\hat{\alpha}_k$ and $\hat{\beta}_k$ is equivalent to taking partial derivatives of the term with θ_k on the right-hand side of (14) with respect to α_k and β_k, and setting them to zero. Substituting (31) for $\log\left(p(x_i|\theta_k)\right)$ in (14) gives

$$\sum_{k=1}^{K}\sum_{i=1}^{M}\left[(\alpha_k-1)\log x_i-\alpha_k\log\beta_k-\log\Gamma(\alpha_k)-\frac{x_i}{\beta_k}\right]w_k^{(i)}, \tag{32}$$

therefore taking $\frac{\partial}{\partial\alpha}$ and $\frac{\partial}{\partial\beta}$ of (32) and setting to them zero gives (34) and (33), respectively.

$$\sum_{i=1}^{M}\left(\frac{x_i}{{\beta_k}^2}-\frac{\alpha_k}{\beta_k}\right)w_k^{(i)}=0, \tag{33}$$

$$\sum_{i=1}^{M}\left(\log x_i-\log\beta_k-\psi(\alpha_k)\right)w_k^{(i)}=0. \tag{34}$$

Solving (33) for β_k gives the closed-form expression

$$\beta_k=\frac{\sum_{i=1}^{M}x_i w_k^{(i)}}{\alpha_k\sum_{i=1}^{M}w_k^{(i)}}=\frac{\bar{x}_i}{\alpha_k}. \tag{35}$$

Substituting for β_k in (34), and simplifying, it results in an update expression for α_k as

$$\psi(\alpha_k)-\log\alpha_k=\log\left(\frac{\sum_{i=1}^{M}x_i w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}\right)-\frac{\sum_{i=1}^{M}\log(x_i)w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}. \tag{36}$$

Therefore, (36) can be rewritten as $\psi(\alpha_k)-\log\alpha_k=\Upsilon$, where Υ is simply a constant. The expression $\psi(\alpha_k)-\log\alpha_k$ is a decreasing function on $(0,\infty)$ since

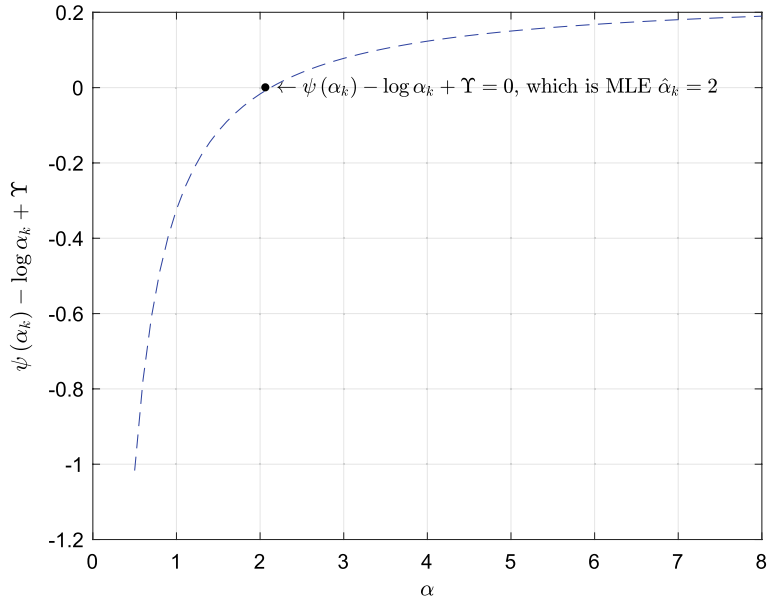

Fig. 3 G_γMM α_k MML estimation

$\lim_{\alpha_k\to\infty} \psi(\alpha_k) - \log\alpha_k = 0$ and $\lim_{\alpha_k\to 0} \psi(\alpha_k) - \log\alpha_k = -\infty$. Since log is a concave function, it follows from Jensen's inequality that

$$\log\left(\frac{\sum_{i=1}^{M} x_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}\right) \geq \frac{\sum_{i=1}^{M} \log(x_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}, \tag{37}$$

the equality in (37) holds if and only if (i.i.f) the samples are mutually equal, which is a case that does not occur in practice. Therefore, $\log\left(\frac{\sum_{i=1}^{M} x_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}\right) - \frac{\sum_{i=1}^{M} \log(x_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} > 0$, and $\psi(\alpha_k) - \log\alpha_k + \Upsilon = 0$, which means that α_k can be found using a bisection root-finding method to find the axis intercept where $\psi(\alpha_k) - \log\alpha_k$ changes sign. To demonstrate this, gamma random variables are simulated using (31) with $\alpha_k = 2$ and $\beta_k = 5$. A plot of $\psi(\alpha_k) - \log\alpha_k + \Upsilon$ is shown in Fig. 3, and the ML estimate of α_k is shown as the point where $\psi(\alpha_k) - \log\alpha_k + \Upsilon = 0$ holds.

3.2.5 IG$_\gamma$MM

For the IG$_\gamma$MM, each x_i is assumed to be generated by one the K inverse gamma densities with shape α_k and scale β_k, which is expressed as

$$p(x_i|\alpha_k, \beta_k) = \frac{{x_i}^{-(\alpha_k+1)}}{{\beta_k}^{\alpha_k}\Gamma(\alpha_k)} \exp\left(-\frac{1}{x_i\beta_k}\right), \quad x_i > 0,\ \alpha_k > 0,\ \beta_k > 0. \tag{38}$$

Substituting (38) for $\log(p(x_i|\theta_k))$ in (14) gives

$$\sum_{k=1}^{K}\sum_{i=1}^{M}\left[-(\alpha_k - 1)\log x_i - \alpha_k \log\beta_k - \log\Gamma(\alpha_k) - \frac{1}{x_i\beta_k}\right] w_k^{(i)}, \tag{39}$$

taking $\frac{\partial}{\partial\beta}$ and $\frac{\partial}{\partial\alpha}$ of (39) and setting them to zero gives (40) and (41), respectively.

$$\sum_{i=1}^{M}\left(\frac{1}{x_i{\beta_k}^2}-\frac{\alpha_k}{\beta_k}\right)w_k^{(i)}=0, \tag{40}$$

$$\sum_{i=1}^{M}\left(-\log x_i-\log\beta_k-\psi(\alpha_k)\right)w_k^{(i)}=0. \tag{41}$$

Let $\frac{1}{x_i}=\varepsilon_i$, then solving (40) for β_k gives the closed-form expression

$$\beta_k=\frac{\sum_{i=1}^{M}\varepsilon_i w_k^{(i)}}{\alpha_k\sum_{i=1}^{M}w_k^{(i)}}=\frac{\bar{\varepsilon}_i}{\alpha_k}. \tag{42}$$

Substituting for β_k in (41), and simplifying, it gives an update expression for α_k as

$$\psi(\alpha_k)-\log\alpha_k=-\log\left(\frac{\sum_{i=1}^{M}\varepsilon_i w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}\right)+\frac{\sum_{i=1}^{M}\log(\varepsilon_i)w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}. \tag{43}$$

Similar to the G$_\gamma$MM, for the IG$_\gamma$MM, (43) can be rewritten as $\psi(\alpha_k)-\log\alpha_k=\Upsilon$, where Υ is simply a constant. $\psi(\alpha_k)-\log\alpha_k$ is a decreasing function on $(0,\infty)$ since $\lim_{\alpha_k\to\infty}\psi(\alpha_k)-\log\alpha_k=0$ and $\lim_{\alpha_k\to 0}\psi(\alpha_k)-\log\alpha_k=-\infty$. Since log is a concave function, it follows from Jensen's inequality that

$$\log\left(\frac{\sum_{i=1}^{M}\varepsilon_i w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}\right)\geq\frac{\sum_{i=1}^{M}\log(\varepsilon_i)w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}, \tag{44}$$

where the equality in (44) holds i.i.f the samples are mutually equal, which is a case that does not occur in practice. Therefore, $\log\left(\frac{\sum_{i=1}^{M}\varepsilon_i w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}\right)-\frac{\sum_{i=1}^{M}\log(\varepsilon_i)w_k^{(i)}}{\sum_{i=1}^{M}w_k^{(i)}}>0$, and $\psi(\alpha_k)-\log\alpha_k+\Upsilon=0$, which means that α_k can be found using a bisection root-finding method to find the axis intercept where $\psi(\alpha_k)-\log\alpha_k$ changes sign. To demonstrate this, inverse gamma random variables are simulated using (38) with $\alpha_k=4$ and $\beta_k=5$. A plot of $\psi(\alpha_k)-\log\alpha_k+\Upsilon$ is shown in Fig. 4, and the ML estimate of α_k is shown as the point where $\psi(\alpha_k)-\log\alpha_k+\Upsilon=0$ holds.

3.2.6 NMM

For the NMM, each x_i is assumed to be generated by one of the K Nakagami densities with shape μ_k and scale Ω_k, which is expressed as

$$p(x_i|\mu_k,\Omega_k)=\frac{2{\mu_k}^{\mu_k}{x_i}^{(2\mu_k-1)}}{\Gamma(\mu_k){\Omega_k}^{\mu_k}}\exp\left(-\frac{{x_i}^2\mu_k}{\Omega_k}\right),\quad x_i>0,\ \Omega_k>0,\ \mu_k\geq\frac{1}{2}. \tag{45}$$

Substituting (45) for $\log\left(p(x_i|\theta_k)\right)$ in (14) gives

$$\sum_{k=1}^{K}\sum_{i=1}^{M}\Big[\log 2+\mu_k\log\mu_k+(2\mu_k-1)\log x_i-\log\Gamma(\mu_k) -\mu_k\log\Omega_k-\frac{{x_i}^2\mu_k}{\Omega_k}\Big]w_k^{(i)}, \tag{46}$$

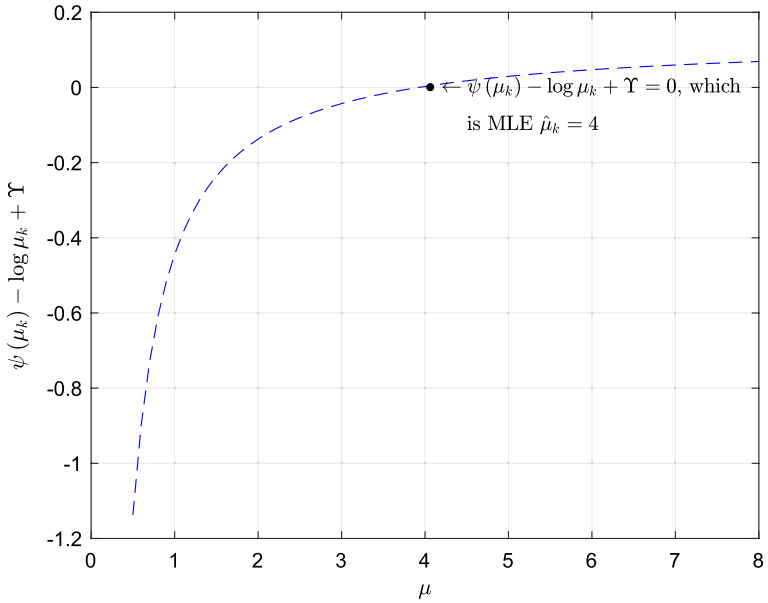

Fig. 4 IG$_\gamma$MM α_k ML estimation

taking $\frac{\partial}{\partial \mu}$ and $\frac{\partial}{\partial \Omega}$ of (46) and setting them to zero gives (48) and (47), respectively.

$$\sum_{i=1}^{M} \left(\frac{{x_i}^2 \mu_k}{{\Omega_k}^2} - \frac{\mu_k}{\Omega_k} \right) w_k^{(i)} = 0, \tag{47}$$

$$\sum_{i=1}^{M} \left(\log \mu_k + 2 \log x_i - \psi(\mu_k) - \log \Omega_k - \frac{{x_i}^2}{\Omega_k} \right) w_k^{(i)} = 0. \tag{48}$$

Let ${x_i}^2 = \upsilon_i$, then solving (47) for Ω_k gives the closed-form expression

$$\Omega_k = \frac{\sum_{i=1}^{M} \upsilon_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} = \bar{\upsilon}_i \tag{49}$$

Substituting for Ω_k in (48), and simplifying, it results in an update expression for μ_k as

$$\psi(\mu_k) - \log \mu_k = \frac{\sum_{i=1}^{M} \log(\upsilon_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} - \log \left(\frac{\sum_{i=1}^{M} \upsilon_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right). \tag{50}$$

$\psi(\mu_k) - \log \mu_k$ is a decreasing function on $(0, \infty)$ since $\lim_{\mu_k \to \infty} \psi(\mu_k) - \log \mu_k = 0$ and $\lim_{\mu_k \to 0} \psi(\mu_k) - \log \mu_k = -\infty$. Since x_i is positive, log is a concave function and it follows from Jensen's inequality that

$$\log \left(\frac{\sum_{i=1}^{M} \upsilon_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right) \geq \frac{\sum_{i=1}^{M} \log(\upsilon_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}. \tag{51}$$

The equality in (51) holds i.i.f the samples are mutually equal. Therefore, $\log \left(\frac{\sum_{i=1}^{M} \upsilon_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right) - \frac{\sum_{i=1}^{M} \log(\upsilon_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} > 0$ and $\psi(\mu_k) - \log \mu_k + \Upsilon = 0$, where Υ is a con-

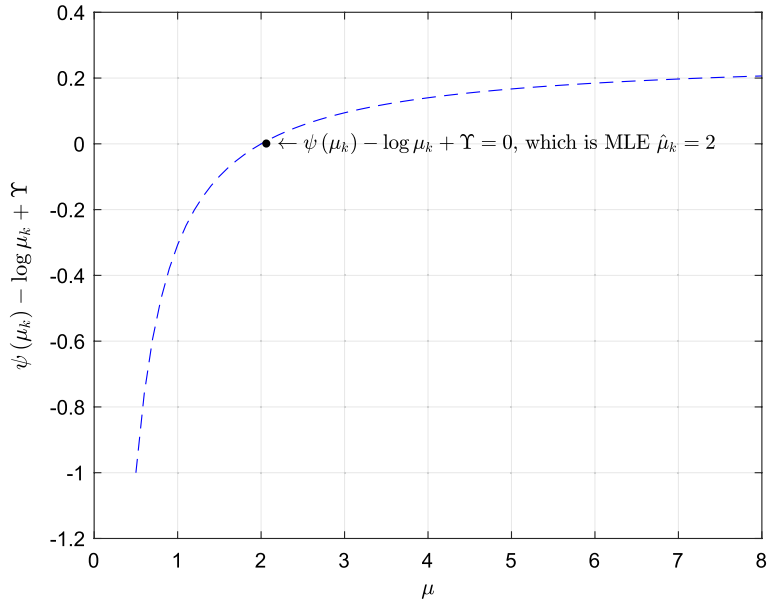

Fig. 5 NMM μ_k ML estimation

stant equivalent to the right-hand side of (50). This means μ_k can be found using a bisection root-finding method to find the axis intercept where $\psi(\mu_k) - \log \mu_k$ changes sign. To demonstrate this, Nakagami random variables are simulated using (45) with $\mu_k = 2$ and $\beta_k = 1.5$. A plot of $\psi(\mu_k) - \log \mu_k + \Upsilon$ for $\mu_k \in [0.5, 8]$ is shown in Fig. 5, and the the ML estimate of μ_k is shown as the point where $\psi(\mu_k) - \log \mu_k + \Upsilon = 0$ holds.

3.2.7 INMM

For the INMM, each x_i is assumed to be generated by one of the K inverse Nakagami densities with shape μ_k and scale Ω_k, which is expressed as

$$p(x_i|\mu_k, \Omega_k) = \frac{2{\mu_k}^{\mu_k} {x_i}^{-(2\mu_k - 1)}}{\Gamma(\mu_k){\Omega_k}^{\mu_k}} \exp\left(-\frac{\mu_k}{\Omega_k {x_i}^2}\right), \quad x_i > 0,\ \Omega_k > 0,\ \mu_k > 0. \tag{52}$$

Substituting (52) for $\log\left(p(x_i|\theta_k)\right)$ in (14) gives

$$\sum_{k=1}^{K}\sum_{i=1}^{M} \left[\log 2 + \mu_k \log \mu_k - (2\mu_k - 1)\log x_i - \log \Gamma(\mu_k) \right.$$
$$\left. -\mu_k \log \Omega_k - \frac{\mu_k}{\Omega_k {x_i}^2}\right] w_k^{(i)}, \tag{53}$$

taking $\frac{\partial}{\partial \mu}$ and $\frac{\partial}{\partial \Omega}$ of (53) and setting them to zero gives (55) and (54), respectively.

$$\sum_{i=1}^{M} \left(\frac{\mu_k}{{\Omega_k}^2 {x_i}^2} - \frac{\mu_k}{\Omega_k}\right) w_k^{(i)} = 0, \tag{54}$$

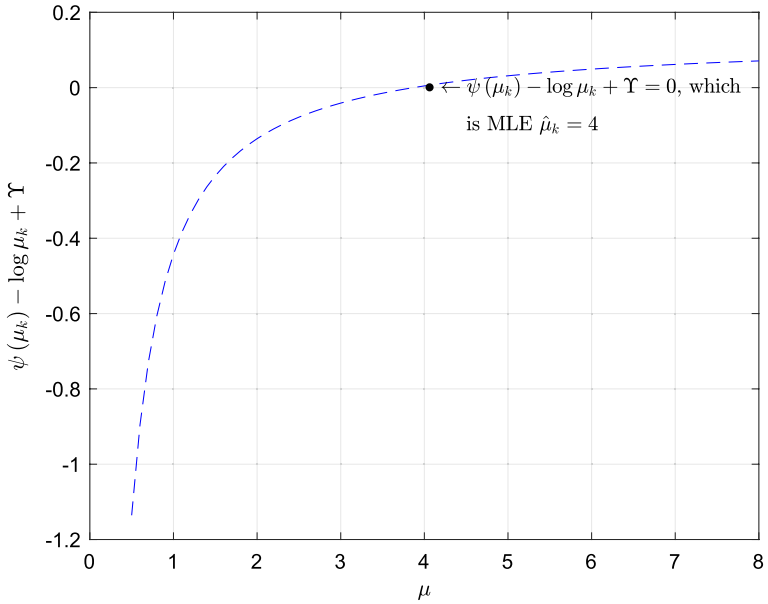

Fig. 6 INMM μ_k ML estimation

$$\sum_{i=1}^{M} \left(\log \mu_k - 2 \log x_i - \psi(\mu_k) - \log \Omega_k - \frac{{x_i}^2}{\Omega_k} \right) w_k^{(i)}. = 0. \tag{55}$$

Let $\frac{1}{{x_i}^2} = \eta_i$, then solving (54) for Ω_k gives the closed-form expression

$$\Omega_k = \frac{\sum_{i=1}^{M} \eta_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} = \bar{\eta}_i. \tag{56}$$

Substituting for Ω_k in (55), and simplifying, it results in an update expression for μ_k as

$$\psi(\mu_k) - \log \mu_k = \frac{\sum_{i=1}^{M} \log(\eta_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} - \log \left(\frac{\sum_{i=1}^{M} \eta_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right). \tag{57}$$

Similarly to the NMM, for the INMM, $\psi(\mu_k) - \log \mu_k$ is a decreasing function on $(0, \infty)$ since $\lim_{\mu_k \to \infty} \psi(\mu_k) - \log \mu_k = 0$ and $\lim_{\mu_k \to 0} \psi(\mu_k) - \log \mu_k = -\infty$. Since x_i is positive, log is a concave function, and it follows from Jensen's inequality that

$$\log \left(\frac{\sum_{i=1}^{M} \eta_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right) \geq \frac{\sum_{i=1}^{M} \log(\eta_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}. \tag{58}$$

The equality in (58) holds i.i.f the samples are mutually equal. Therefore, $\log \left(\frac{\sum_{i=1}^{M} \upsilon_i w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} \right) - \frac{\sum_{i=1}^{M} \log(\upsilon_i) w_k^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}} > 0$ and $\psi(\mu_k) - \log \mu_k + \Upsilon = 0$, where Υ is a constant equivalent to the right-hand side of (57). This means μ_k can be found using a bisection root-finding method to find the axis intercept where $\psi(\mu_k) - \log \mu_k$ changes sign. Inverse Nakagami random variables are simulated using (52) with $\mu_k = 4$ and $\beta_k = 1.5$. A plot of $\psi(\mu_k) - \log \mu_k + \Upsilon$ for $\mu_k \in [0.5, 8]$ is shown in Fig. 6, and the ML estimate of μ_k is shown as the point where $\psi(\mu_k) - \log \mu_k + \Upsilon = 0$ holds.

3.3 Feasible range of clusters

Both distance-based and model-based clustering solutions require a predefined range of clusters such that $k \in \{1, \ldots, K\}$. A method is introduced for estimating the cluster range from the MPC dataset in a manner that guarantees that the optimal number of clusters, K_{opt}, is within the specific range. The feasible range $\mathbf{R} = \{K_1, \ldots, K_N\} \in [K_{\min}, K_{\max}]$ is obtained by algorithm 1. The algorithm takes as input the magnitude-delay dataset of the MPCs, which is estimated from the channel response. $K_{\max}$ is obtained by comparing the magnitudes x_{i-1} with x_i, if $x_{i-1} > x_i$, this means the magnitudes are decreasing with τ. However, if $x_{i-1} < x_i$, this could be the start of another cluster, and $K_{\max}$ is incremented by 1. $K_{\min}$ is obtained by first sorting the indexed MPC magnitudes in decreasing order, then initialising dIx to the first element of the sorted indexes Ix, which is the largest magnitude. Then, if $dIx < Ix_i$, $dIx = Ix_i$ and $K_{\min}$ is incremented by 1.

Algorithm 1 Cluster range estimation

```
procedure ESTIMATE K_min AND K_max FROM α_{m,l} AND τ_{m,l} MPC PARAMETERS
    input the α and τ dataset
    K_max = 0
    for i = 2 : length of α do
        if |α_{i−1}| > |α_i| then
            K_max = K_max + 0
        else
            K_max = K_max + 1
        end if
    end for
    sort the indexed α in decreasing order and store
    the indexes in variable Ix, initialise dIx to
    the first element of Ix
    K_min = 0
    for i = 2 : length of α do
        if Ix(i) > dIx then
            dIx = Ix(i) >
            K_min = K_min + 1
        else
            K_min = K_min + 0
        end if
    end for
end procedure
```

3.4 Corrected Akaike information criterion (AIC_c)

Once the feasible range of clusters has been obtained, it is necessary to find the best cluster partition solution K_{opt} for model-based clustering. For all candidate models and for each $k \in \{1, \ldots, K\}$, AIC_c is calculated as

$$\text{AIC}_\text{c} = -2 \log\left(\mathcal{L}\left(\hat{\theta}_{\text{ML}}|x\right)\right) + 2d + \frac{2d(d+1)}{M-d-1}, \tag{59}$$

where $\mathcal{L}\left(\hat{\theta}_{\text{ML}}|x\right)$ is the value of the maximised log-likelihood, which correspond to the computed log-likelihood function using the ML estimates, d is the number of parameters in the model and M is the number of samples. The first term of (59) indicates the overall fit of the model to the dataset and tends to decrease with d. The second term of (59) penalises the model for increased d to ensure the best model has the least number of parameters. The third term of (59) is a bias term to correct the AIC when M is small.

However, when M is large enough, the bias term becomes negligible. Therefore, K_{opt} is the cluster partition with the smallest AI_c value.

3.5 Cluster validation

The effectiveness of the model-based clustering solution over the distance-based solution is demonstrated through CH and DB CVIs. These CVIs are well-suited for evaluating the separation between clusters and compactness within a cluster. Considering M MPCs in cluster K, the CH index is given as

$$\text{CH}_k = \frac{(M-K)\sum_{k=1}^{K} M_k \cdot \text{MCD}(c_k, \bar{c})^2}{(K-1)\sum_{k=1}^{K}\sum_{i\in c_k} \text{MCD}(x_i, c_k)^2}, \tag{60}$$

where M_k is the number of MPCs in the Kth cluster and $\bar{c}$ is the global centroid computed as the average of all the MPCs. The summation in the numerator of (60) evaluates the separation between clusters, and the summation in the denominator evaluates the compactness within the kth cluster. After computing CH_k for $K \in [K_{\min}, K_{\max}]$, K_{opt} is obtained as

$$K_{\text{opt}} = \arg\max_{K} \{\text{CH}_K\} \tag{61}$$

Considering M MPCs in cluster K, the DB index is given as

$$\text{DB}_K = \frac{1}{K}\sum_{k=1}^{K} R_k, \tag{62}$$

where R_K is given as

$$R_K = \max_{\substack{j=1,\ldots,K \\ j\neq i}} \left\{ \frac{S_i + S_j}{g_{ij}} \right\}, \tag{63}$$

where $S_k = (1/M_k)\sum_{i\in c_k} \text{MCD}(x_i, c_k)$ denotes the cluster compactness and $d_{ij} = \text{MCD}(c_i, c_j)$ denotes the cluster separation. After computing DB_k for $K \in [K_{\min}, K_{\max}]$, K_{opt} is obtained as

$$K_{\text{opt}} = \arg\min_{K} \{\text{DB}_K\}. \tag{64}$$

4 Measurements and results

4.1 Channel measurements and parameter extraction

A PLC channel was constructed using an H05RR-F 3-Core (0.75 mm^2) cable-type, which is typically found in an indoor LV power network. Figure 7 shows how the experimental LV PLC channel was constructed within a residential apartment with a total area of 95 m^2. The channel was unknown since some electrical outlets were left open, and other outlets had common household appliances connected to them. The absolute distances between the outlets were unknown. Capacitive coupling circuits were interfaced to each outlet on the channel. A coupling circuit was used at

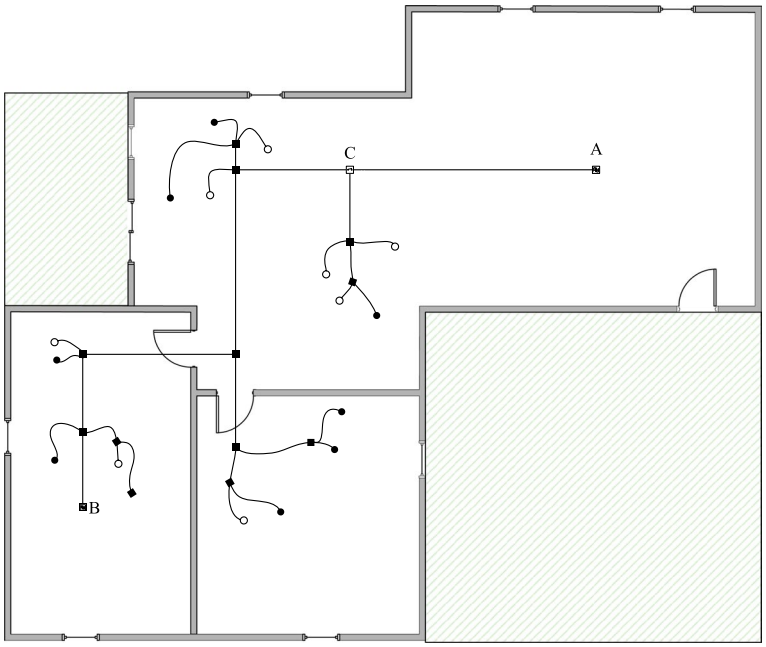

Fig. 7 Experimental LV PLC channel setup

branch C, and this is where the channel was fed with the 230V 50Hz power signal. The authors highlight the following about the constructed channel: (1) each terminal point only has a single node, whether open or with load connected; (2) there are no connections between termination points; and (3) the channel has a radial topology with no loops. At point A, a BladeRF x40 software-defined radio (SDR) is used as a transmitter. Another BladeRF x40 SDR is used as a receiver at point B. GNU Radio platform was used to control the SDRs and capture the channel response. Off-line processing was done in MATLAB.

On the transmitter side, a sounding signal in the form $g(t) = \sum_{q=1}^{N_q} g_q w(t - qT_w)$ was continuously transmitted to excite the channel. g_t is a maximum length sequence (MLS) of length $N_q = 2^L - 1$. L was chosen as 10, therefore $N_q = 1024$. Initially, $\{g_q, \ldots, g_{N_q}\}$ is generated such that $g_q \in \{1, -1\}$; however, for transmitting, it is transformed such that $g_q \in \{1, 0\}$. $w(.)$ is a rectangular pulse shape, and T_w is the pulse duration. The sampling frequency of the BladeRF SDR was set to $f_s = 30$ MHz. Therefore, $T_w = 33.33$ ns and the sequence duration $T_q = 34.13\,\mu$ s. Sounding sequence $g(t)$ was transmitted into the channel, and $y(t)$ in (1) represents the I/Q samples that are captured and stored for off-line processing. The rest of the channel sounder parameters are summarised in Table 1.

The SAGE estimation algorithm was used to obtain the $\alpha_{m,l}$ and $\tau_{m,l}$ parameters of the MPCs from $y(t)$. The SAGE algorithm requires the number of MPCs M_l as input, since the channel was unknown, M_l was unknown and was estimated from $y(t)$. An AIC estimator was used to obtain M_l from $y(t)$ by computing the covariance matrix of $y(t)$ to obtain the eigenvalues that were used to estimate M_l. Eigenvalue-based estimators such as AIC [28], MDL [29] and random matrix theory (RMT) [30] are typically used to estimate the number of signal components in a received noisy channel response and source enumeration in array processing.

Table 1 Channel sounder parameters

Parameter	Value
Carrier frequency (MHz)	415
Bandwidth (MHz)	18
Transmit power (dBm)	19.5
Sampling rate (MSps)	30
MLS sequence length	1024
MLS sequence duration (μs)	34.13

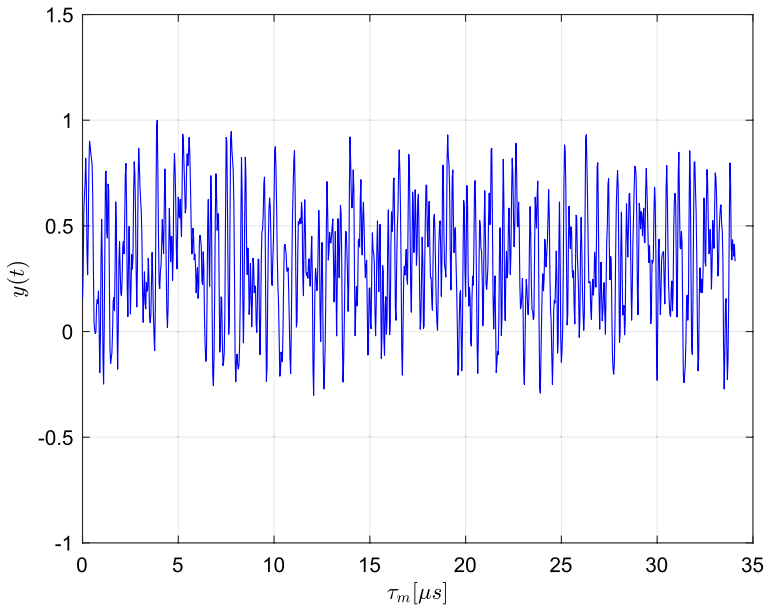

Fig. 8 $y(t)$ channel response to $g(t)$

A single snapshot of $y(t)$ is given in Fig. 8. This is the input to the eigenvalue AIC estimator that outputs $M_l = 56$. Then, M_l, $y(t)$ and parameters of Table 1 are inputs of the SAGE algorithm, which extracts magnitudes and delays. The channel delay profile is shown in Fig. 9. The largest component in the delay profile of Fig. 9 is taken as the 0th, $\tau(0)$, component. This component represents the direct path from transmitter to receiver, and is the first component of the first cluster. In this way, one is not interested in the absolute delays, but rather, in the relative delays of the MPCs for clustering.

4.2 AIC$_c$ model selection

Once the MPC parameters are extracted from the measurements, the feasible range of clusters, $\mathbf{R}$, is estimated using algorithm 1. This gives $K_{\min} = 3$ and $K_{\max} = 21$, which means $K \in [3, 21]$. For $K \in [3, 21]$, the extracted MPC magnitude-delay dataset is fitted to each FMM using the procedures described in Sect. 3.2. Each FMM fit procedure computes 100 iterations between the E-step and the M-step in the EM algorithm. Convergence is reached when $\left\|\Theta^{t+1} - \Theta^{t}\right\| < 10^{-5}$. Parameter estimates of each component PDF for each FMM are initialised using a method of moments (MoM) estimator, and the priors are estimated uniformly. The initial indexing of each

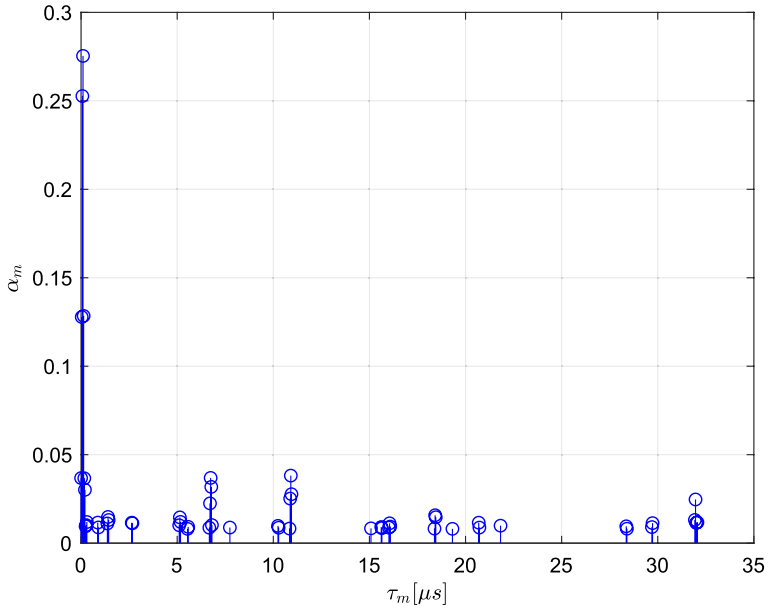

Fig. 9 Channel delay profile

Table 2 AIC values estimated for $K \in [3, 21]$

Mixture model	G_γMM	IG_γMM	NMM	INMM	GMM	RMM	IGMM
$K = 3$	− 334.1857	− 350.9654	− 347.9581	**− 351.5738**	− 313.9865	− 295.5700	− 350.3486
$K = 4$	− 333.2412	− 355.4684	− 354.3063	**− 355.5885**	− 330.8767	− 290.2671	− 355.3550
$K = 5$	− 329.9266	− 353.0241	− 363.4997	**− 364.2100**	− 358.1729	− 284.4337	− 363.5221
$K = 6$	− 331.4286	− 361.1313	**− 368.5970**	− 365.8892	− 367.5014	− 277.9864	− 368.2715
$K = 7$	− 319.8902	− 354.8533	− 365.4429	**− 372.7487**	− 361.9111	− 270.8226	− 361.1077
$K_{opt} = 8$	− 317.4058	− 349.3820	− 357.3971	**− 383.5133**	− 336.5550	− 262.8161	− 357.4874
$K = 9$	− 309.3846	− 349.6467	− 362.3849	**− 370.5133**	− 343.3241	− 253.8087	− 356.7042
$K = 10$	− 301.5775	− 369.4797	− 371.7292	− 367.5000	− 345.0854	− 243.6004	**− 372.9758**
$K = 11$	− 290.2804	− 330.1499	− 339.8575	**− 367.3161**	− 321.4491	− 231.9338	− 334.8292
$K = 12$	− 286.1923	− 348.3962	− 328.3997	− 350.1985	− 314.5583	− 218.4722	**− 361.4057**
$K = 13$	− 267.6066	− 304.1411	− 303.4176	**− 331.8072**	− 289.5755	− 202.7671	− 337.1425
$K = 14$	− 253.4266	− 305.3975	− 286.5254	**− 345.9448**	− 285.6640	− 184.2065	− 327.0185
$K = 15$	− 230.5696	− 294.2617	− 275.4093	**− 325.6961**	− 268.6821	− 161.9338	− 305.9090
$K = 16$	− 203.3850	− 271.8704	− 234.0650	**− 300.8839**	− 282.6890	− 134.7116	− 277.5505
$K = 17$	− 169.3280	− 233.0117	− 206.4532	**− 257.7724**	− 277.4756	− 100.6838	− 243.4957
$K = 18$	− 125.5837	− 177.7226	**− 200.9584**	− 195.1257	− 158.4187	− 56.9338	− 206.1883
$K = 19$	− 69.7917	− 129.6232	− 118.0975	**− 162.6962**	− 121.3401	1.3995	− 143.2777
$K = 20$	13.4495	− 55.7850	− 18.7447	**− 74.0224**	− 100.1851	83.0662	− 66.1916
$K = 21$	103.0252	− 1.1628	61.8242	− 15.5922	**− 53.8652**	182.5822	− 11.0423

The best-fit model for each Kth cluster index was shown in bold

MPC, for the Kth cluster, is done using the k-means method. Once components have been indexed, MPC clustering is refined with FMMs.

Table 2 lists AIC_c values computed for each FMM for $K \in [3, 21]$, and Fig. 10 shows a plot of the AIC_c values. Overall, the INMM can be observed to obtain the

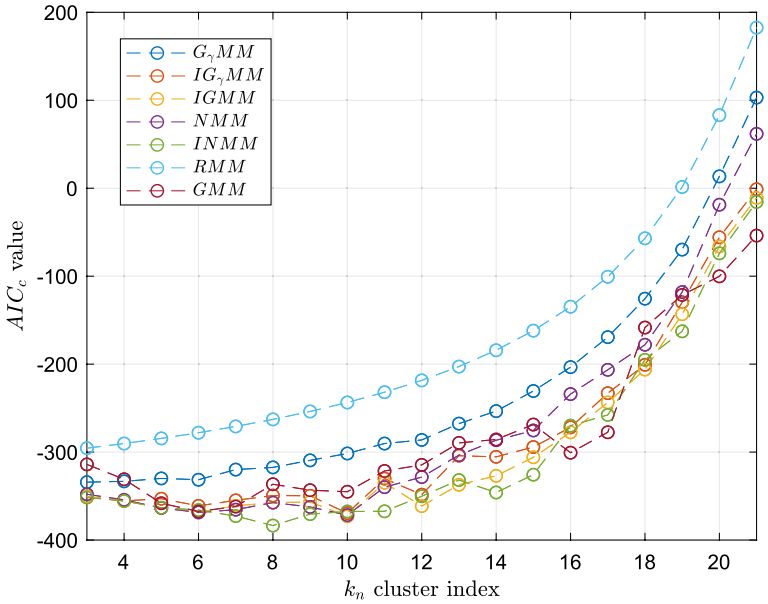

Fig. 10 AIC_c results for $K \in [3, 21]$

best clustering solution because it consistently obtains the minimum AIC_c value. As described in Sect. 3.4, the lowest AIC_c value indicates the best clustering solution for K clusters. The absolute value of AIC_c is not very important because AIC_c is used to find the smallest value from all candidate models. This value can be positive or negative. The lowest AIC_c value for each K is listed in bold in Table 2. For $K = 5$ and $K = 19$, NMM outperforms the INMM, and for $K = 10$ and $K = 12$, the IGMM outperforms the INMM. The RMM is observed to have the worst performance since it consistently obtains the highest AIC_c value for all $K \in [3, 21]$. This can be attributed to the Rayleigh model having only a shape parameter. The rest of the FMMs are two-parameter models that capture the shape and scale of the distribution. Such models are therefore more effective in modelling the positively skewed distribution of the channel delay profile.

Optimal clustering is obtained with the INMM for $K_{\text{opt}} = 8$, which has $AIC_c = -383.5133$. This means that the channel delay profile in Fig. 9 is best described as having eight clusters of MPCs. Figure 11 shows the optimal cluster partitions for the MPCs of channel delay profile that are obtained using the INMM. The first four clusters overlap with only a few MPCs. This can be attributed to the propagation paths of the MPCs of the overlapping clusters having similar propagation delays. The last four cluster are distinct without any overlap. This can be attributed to the propagation paths of the MPCs in these clusters having distinct and long propagation delays. The AIC_c values of the NMM and INMM show that the two FMMs obtain comparable clustering performance because they are related by a simple inverse of the random variable. If X is a random variable generated by a Nakagami distribution, then $Y = 1/X$ is a random variable generated by an inverse Nakagami distribution. In a similar manner, the inverse gamma distribution is also related to the gamma distribution by an inverse of the random variable.

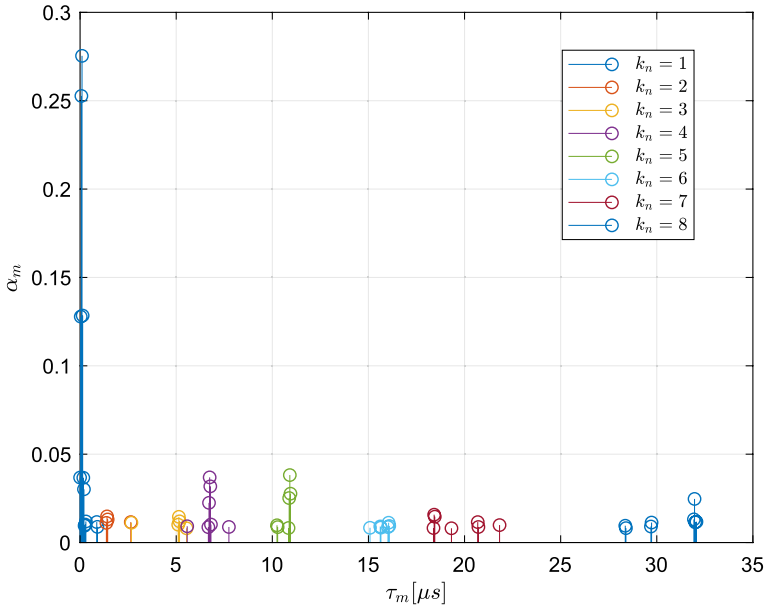

Fig. 11 Optimal cluster partitions of MPCs using the INMM

4.3 Cluster validation

Cluster index validation was done by computing the CH index and DB index for $K \in [3, 21]$. In simulation or when a channel is known, the CH index will give a maximum value, and the DB index will give a minimum value for the correct number of clusters. However, in this case, the channel is unknown. Therefore, CVIs are used to show that the model-based clustering solution produces more favourable results than distance-based methods.

As described in Sect. 3.5, the CH and DB indexes evaluate within-cluster compactness and between-cluster separation. Performing clustering for $K \in [3, 21]$ means that for some low values of K, all clusters would be indexed, i.e. if $K = 4$, each MPC would be indexed with i such that $i \in [1, 4]$. In this case, the within-cluster compactness improves, while the between-cluster separation deteriorates. For higher values of K, there would be some empty clusters, i.e. if $K = 14$, MPCs would be indexed with i such that $i = 2, 4, 5, 8, 9, 10, 11, 12, 13$. This means some clusters would have no components allocated. In such a case, the within-cluster compactness deteriorates, while the between-cluster separation improves. Model-based clustering solutions can effectively optimise cluster compactness and separation by allowing a component to belong to multiple clusters at the same time, but with different probabilities.

To illustrate the effectiveness of model-based clustering, Fig. 12 shows a plot of CH index values computed for each FMM clustering solution and clustering by k-means. A high CH value corresponds to the most favourable clustering solution. It is clear that k-means obtains the lowest CH values for $K \in [5, 21]$, and even for $K = 3$ and $K = 4$, k-means is outperformed by five FMMs with higher CH index values. The highest CH index is obtained by the INMM for $K = 8$. Figure 13 shows a plot of DB index values computed for each FMM clustering solution and clustering by k-means. Using the DB index, a low value corresponds to the most favourable clustering

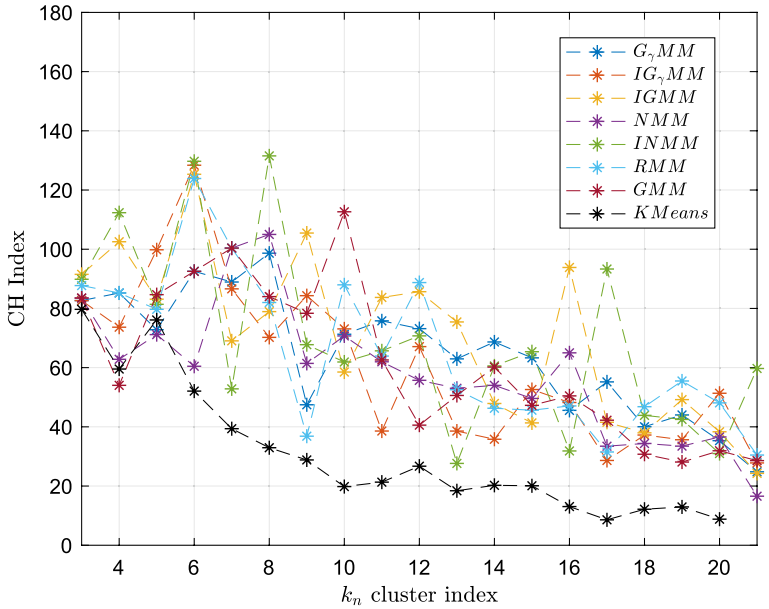

Fig. 12 CH index results for $K \in [3, 21]$

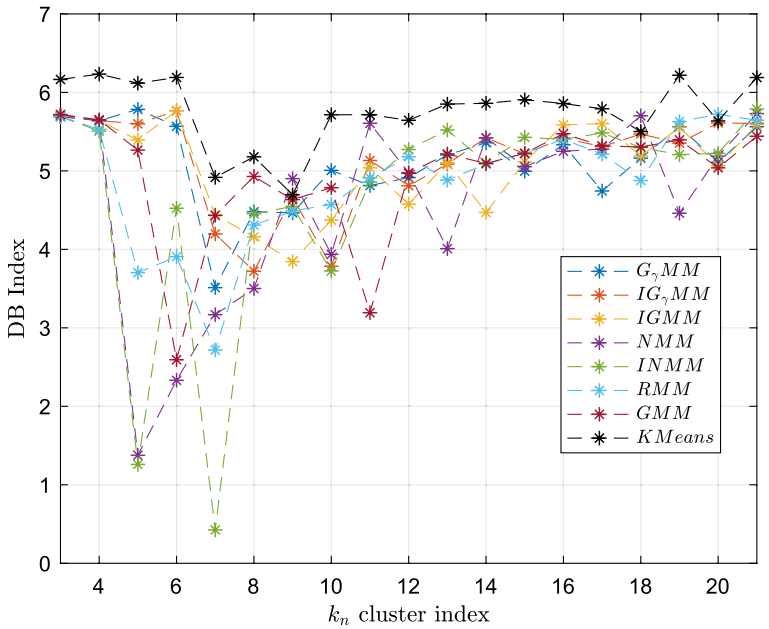

Fig. 13 DB index results for $K \in [3, 21]$

solution. It is clear that k-means obtains the highest DB values for $K \in [3, 21]$, with the lowest DB index obtained for the INMM for $K = 7$.

These results show that even though the PLC channel is unknown, CH and DB indexes are comparable in terms of K_{opt}. Therefore, $K_{\text{opt}} \in [7, 8]$ corresponds to the most within-cluster compactness of MPCs and the most between-cluster separation in the delay domain. Cluster compactness and separation characteristics are thus best captured using a model-based clustering solution. A keen observation made is that the optimal number

of clusters K_{opt} does not necessarily equal the number of branch-connection clusters. Typically, MPCs reflected from a branch-connection cluster would be finite, weak and exponentially decaying. Moreover, channel impairments such as cable skin-effect, cable branching and connected loads would attenuate the already weak signal components if they were to propagate into another branch-connection cluster. However, this observation suggests that some strong signal components reflected from one branch-connection cluster would propagate into another branch-connection cluster, resulting in a secondary MPC cluster. Further investigation of indoor LV PLC channels is needed to study the effect and contribution of secondary MPC clusters to the overall channel impulse response. Current prevailing cluster-based impulse response models such as the SV model require that the distribution of MPC time instances and cluster arrival times to be conditioned on previous time instances. As demonstrated in this study, an MPC cluster can be treated as an independent component distribution of the mixture distribution, and is not conditioned on any previous MPC time instances. Therefore, reconstruction of the channel impulse response using the optimal FMM for an optimal number of clusters warrants further investigation of the model-based clustering methodology.

5 Conclusions

This paper addressed the multipath clustering problem in PLC channels using both distance-based and model-based methods. The problem formulation considers an indoor LV power network with branch-connection clusters, resulting in clusters of MPCs for a propagating signal. A measurement campaign of a constructed but unknown PLC channel with branch-connection clusters was conducted using an MLS channel sounding method. An eigenvalue AIC estimator was used to find the number of MPCs from the channel response. The SAGE algorithm was then used to extract the MPC magnitude-delay parameters from the channel response. A novel method of estimating the feasible range of clusters from the extracted MPCs was introduced. The feasible range was used in both distance-based and model-based clustering solutions. An ML approach was used for fitting the FMMs to the extracted MPCs. This proved to be efficient in estimating parameters for both closed-form and update expression solutions. The AIC_{c} results show that the optimal number of clusters was obtained using the finite INMM. CH and DB CVIs' results also show that the finite INMM obtained the best performance in terms of within-cluster compactness and between-cluster separation. Moreover, CH and DB CVIs' results show that even though the channel was unknown, the distance-based clustering solution obtained the worst performance over the feasible range of clusters. Optimal cluster partitions are obtained using the model-based clustering solution. For each CVI, the optimal cluster partitions are comparable in terms of within-cluster compactness and between-cluster separation in the delay domain. However, reconstruction of the channel impulse response using the optimal FMM for an optimal number of clusters warrants further investigation of the model-based clustering methodology.

Abbreviations

AIC	Akaike information criterion
AOA	Angle of arrival
AOD	Angle of departure
AWGN	Additive white Gaussian noise
CH	Calinski–Harabasz

CVI	Cluster validation index
DB	Davies–Bouldin
DCM	Direction channel model
EM	Expectation–maximisation
EOA	Elevation angle of arrival
EOD	Elevation angle of departure
FMM	Finite-mixture model
GBSM	Geometry-based stochastic model
GMM	Gaussian mixture model
G_γMM	Gamma mixture model
IGMM	Inverse Gaussian mixture model
IG_γMM	Inverse gamma mixture model
INMM	Inverse Nakagami mixture model
JSED	Joint squared Euclidean distance
kPM	KPowerMeans
LV	Low-voltage
MCD	Multipath component distance
ML	Maximum likelihood
MLS	Maximum length sequence
MoM	Method of moments
MPC	Multipath propagation component
NMM	Nakagami mixture model
PDF	Probability density function
PDP	Power delay profile
PLC	Powerline communication
PSD	Power spectral density
RMM	Rayleigh mixture model
RMT	Random matrix theory
SAGE	Space-alternating generalised expectation maximisation
SCM-MIMO	Spatial channel model for multiple-input multiple-output
SDR	Software defined radio
SED	Squared Euclidean distance
SV	Saleh–Valenzuela
WB	Wideband

Acknowledgements
Not applicable.

Author contributions
KM conceived of the presented idea and wrote the manuscript. HM provided valuable advice on the methodology and checked the manuscript. Both authors read and approved the final manuscript.

Funding
Not applicable

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request

Declarations

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Received: 23 May 2023 Accepted: 21 September 2023
Published online: 03 October 2023

References

1. P.M. Shankar, *Fading and Shadowing in Wireless Systems*, 2nd edn. (Springer, New York, 2017)
2. H. Asplund, A.F. Molisch, M. Steinbauer, N.B. Mehta, Clustering of scatterers in mobile radio channels—evaluation and modeling in the COST259 directional channel model, in *IEEE International Conference on Communications*, vol. 2 (2002), pp. 901–905. https://doi.org/10.1109/ICC.2002.996986
3. A.A.M. Saleh, R. Valenzuela, A statistical model for indoor multipath propagation. IEEE J. Sel. Areas Commun. **5**(2), 128–137 (1987). https://doi.org/10.1109/JSAC.1987.1146527

4. P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, G.D. Laurenson, A.F. Matz, C. Molisch, H.. Özcelik. Oestges, Survey of channel and radio propagation models for wireless MIMO systems. EURASIP J. Wirel. Commun. Netw. (2007). https://doi.org/10.1155/2007/24595
5. N. Czink, P. Cera, J. Salo, E. Bonek, J.-P. Nuutinen, J. Ylitalo, Automatic clustering of MIMO channel parameters using the multi-path component distance measure, in *International Symposium on Wireless Personal Multimedia Communications* (2005). https://doi.org/10.1109/ICC.2002.996986
6. N. Czink, C. Mecklenbrauker, A novel automatic cluster tracking algorithm, in *International Symposium on Personal, Indoor and Mobile Radio Communications* (2006), pp. 1–5. https://doi.org/10.1109/PIMRC.2006.254347
7. J. Laurila, K. Kalliola, M. Toeltsch, K. Hugl, P. Vainikainen, E. Bonek, Wideband 3D characterization of mobile radio channels in urban environment. IEEE Trans. Antennas Propag. **50**(2), 233–243 (2002). https://doi.org/10.1109/8.998000
8. L. Vuokko, P. Vainikainen, J. Takada, Clusters extracted from measured propagation channels in macrocellular environments. IEEE Trans. Antennas Propag. **53**(12), 4089–4098 (2005). https://doi.org/10.1109/TAP.2005.859763
9. K. Yu, Q. Li, M. Ho, Measurement investigation of tap and cluster angular spreads at 5.2 GHz. IEEE Trans. Antennas Propag. **53**(7), 2156–2160 (2005). https://doi.org/10.1109/TAP.2005.850721
10. N. Czink, P. Cera, J. Salo, E. Bonek, J.-P. Nuutinen, J. Ylitalo, A framework for automatic clustering of parametric MIMO channel data including path powers, in *IEEE Vehicular Technology Conference* (2006), pp. 1–5. https://doi.org/10.1109/VTCF.2006.35
11. C. Schneider, M. Bauer, M. Narandzic, W.A.T. Kotterman, R.S. Thoma, Clustering of MIMO channel parameters—performance comparison, in *IEEE Vehicular Technology Conference* (2009), pp. 1–5. https://doi.org/10.1109/VETECS.2009.5073445
12. J.C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J. Cybern. **3**(3), 32–57 (1973). https://doi.org/10.1080/01969727308546046
13. D.L. Davies, D.W. Bouldin, A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI–1**(2), 224–227 (1979). https://doi.org/10.1109/TPAMI.1979.4766909
14. G.W. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**(2), 159–179 (1985). https://doi.org/10.1109/VTCF.2006.35
15. O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, An extensive comparative study of cluster validity indices. Pattern Recognit. **46**(1), 243–256 (2013). https://doi.org/10.1016/j.patcog.2012.07.021
16. C.M. Bishop, *Pattern Recognition and Machine Learning*, 1st edn. (Springer, New York, 2006)
17. B.H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, K. Ingeman Pedersen, Channel parameter estimation in mobile radio environments using the sage algorithm. IEEE J. Sel. Areas Commun. **17**(3), 434–450 (1999). https://doi.org/10.1109/49.753729
18. M. Zimmermann, K. Dostert, A multipath for the powerline channel. IEEE Trans. Commun. **50**(4), 553–559 (2002)
19. A.M. Tonello, F. Versolatto, A. Pittolo, In-home power line communication channel: statistical characterization. IEEE Trans. Commun. **62**(6), 2096–2106 (2014). https://doi.org/10.1109/TCOMM.2014.2317790
20. S. Galli, A novel approach to the statistical modeling of wireline channels. IEEE Trans. Commun. **59**(5), 1332–1345 (2011). https://doi.org/10.1109/TCOMM.2011.031611.090692. arXiv:1101.1915v1
21. Y. Li, J. Zhang, Z. Ma, Clustering in wireless propagation channel with a statistics-based framework, in *IEEE Wireless Communications and Networking Conference* (2018), pp. 1–6. https://doi.org/10.1109/WCNC.2018.8377218
22. Y. Li, J. Zhang, Z. Ma, Y. Zhang, Clustering analysis in the wireless propagation channel with a variational Gaussian mixture model. IEEE Trans. Big Data **6**(2), 223–232 (2020). https://doi.org/10.1109/TBDATA.2018.2840696
23. A.M. Tonello, F. Versolatto, A. Pittolo, In-home power line communication channel: statistical characterization. IEEE Trans. Commun. **62**(6), 2096–2106 (2014). https://doi.org/10.1109/TCOMM.2014.2317790
24. S. Galli, T.C. Banwell, A deterministic frequency-domain model for the indoor power line transfer function. IEEE J. Sel. Areas Commun. **24**(7), 1304–1315 (2014). https://doi.org/10.1109/JSAC.2006.874428.0409053
25. J.A. Corchado, J.A. Cortes, F.J. Canete, L. Diez, An MTL-based channel model for indoor broadband MIMO power line communications. IEEE J. Sel. Areas Commun. **34**(7), 2045–2055 (2016). https://doi.org/10.1109/JSAC.2016.2566178
26. M. Zimmermann, K. Dostert, Analysis and modeling of impulsive noise in broad-band powerline communications. IEEE Trans. Electromagn. Compat. **44**(1), 249–258 (2002). https://doi.org/10.1109/15.990732
27. K.P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, 1st edn. (MIT Press, Cambridge, 2012)
28. E. Fishler, H.V. Poor, Estimation of the number of sources in unbalanced arrays via information theoretic criteria. IEEE Trans. Signal Process. **53**(9), 3543–3553 (2005). https://doi.org/10.1109/TSP.2005.853099
29. L. Huang, H.C. So, Source enumeration via MDL criterion based on linear shrinkage estimation of noise subspace covariance matrix. IEEE Trans. Signal Process. **61**(19), 4806–4821 (2013). https://doi.org/10.1109/TSP.2013.2273198
30. S. Kritchman, B. Nadler, Non-parametric detection of the number of signals: hypothesis testing and random matrix theory. IEEE Trans. Signal Process. **57**(10), 3930–3941 (2009). https://doi.org/10.1109/TSP.2009.2022897

Publisher's Note