EM-Based Channel Estimation from Crowd-Sourced RSSI Samples Corrupted by Noise and Interference

Silvija Kokalj-Filipovic* and Larry Greenstein
WINLAB, Rutgers University
{skokalj,ljg}@winlab.rutgers.edu

Abstract—We propose a method for estimating channel parameters from RSSI measurements and the lost packet count, which can work in the presence of losses due to both interference and signal attenuation below the noise floor. This is especially important in the wireless networks, such as vehicular, where propagation model changes with the density of nodes. The method is based on Stochastic Expectation Maximization, where the received data is modeled as a mixture of distributions (no/low interference and strong interference), incomplete (censored) due to packet losses. The PDFs in the mixture are Gamma, according to the commonly accepted model for wireless signal and interference power. This approach leverages the loss count as additional information, hence outperforming maximum likelihood estimation, which does not use this information (ML-), for a small number of received RSSI samples. Hence, it allows inexpensive on-line channel estimation from ad-hoc collected data. The method also outperforms ML- on censored data mixtures, as ML- assumes that samples are from a single-mode PDF.†

I. INTRODUCTION

For various reasons (such as participatory RF sensing in order to develop low-cost RF maps [1], or for calibrating the channel in order to reproduce field trials in a simulator), wireless systems often collect signal strength data on the fly, i.e., in the course of actual operation. Such data is often collected in the form of paired values of Tx-Rx distance and the received signal strength indication (RSSI), which can be thought of (within a known additive constant) as the received power in dBm [2]. RSSI is measured on a per-packet basis. If there is too much noise and/or interference for a given measurement, the packet can be lost in which case only the failure indication is recorded (indirectly, e.g., through packet sequencing). The data reduction challenge is to reconstruct, from the collection of recorded RSSI values and packets tagged as lost, the probability density function (PDF) of the received signal. With the PDF thus estimated, the analyst can accurately model the propagation in the environment (e.g., path loss vs. distance), and also model interference effects for a given scenario (e.g., geometry, spatial density of both active and inactive Tx-s, etc.) The widespread adoption of Nakagami PDFs for modeling radio links is justified by the abundant analysis of empirical data [3], [4]. When we refer to the Nakagami PDF, it implies the signal amplitude; the corresponding power is Gamma-distributed, with the same scale parameter m and shape parameter Ω, and the dB power (hence, RSSI) can be thought of as log-Gamma. Note that Nakagami with m=1 corresponds to the Rayleigh distribution.

However, the problem of estimating parameters of this PDF based on packet data collected over time periods of practical interest (the shorter the better) remains challenging. The reason is a high amount of lost (censored) samples caused by interference and low SNR due to fading or distance-based attenuation. As interference is intermittent, there are two broad classes of RSSI data points, namely, those with no (or low) interference, and those with enough interference to result in a significantly modified statistical model (different PDF). Note that maximum-likelihood (ML) [5], the typically best approach for single statistical model, does not offer a closed form solution for data mixtures with loss counts. To derive parameters of PDFs featured in a censored mixture of two random variables (RVs), representing samples with no/low interference, and with strong interference, we propose the use of Stochastic Expectation-Maximization (SEM) [6] estimators. In addition, our approach leverages the loss count as additional information to improve the estimation accuracy for a given number of samples. We introduce notation ML- to denote the ML that utilizes a single-mode PDF assumption and only received samples. In this paper, we demonstrate that our approach performs better than ML- in the presence of interference, because it starts with an assumption of two components (dual mixture) and because it uses the loss count as side information. It also outperforms ML- in cases without interference, if the number of received samples is small, which is frequently the case in on-line estimation tasks.

The organization of the paper is as follows: in Section II we briefly describe the system model based on an example, while in Section III we introduce basic algorithmic elements; in Section IV we present the algorithm used in our approach; we evaluate our model on both simulated and empirical data, and discuss the results in Section V. In the last section we conclude and address future work.

II. SYSTEM MODEL AND MOTIVATIONAL EXAMPLE

We refer to no/low interference samples as signal (or 1st) component, and to strong interference samples as interference (or 2nd) component. We propose to have both signal and interference components in the mixture modeled by the same
family of PDFs, i.e., Gamma. Properly parameterized Gamma PDFs (GPDFs) are widely used to model small-scale fading, to approximate the product of the small-scale and lognormal fading distribution, and to approximate the interference power [7]. Our claim that interference samples deserve to be modeled by a 2nd component is evident in Fig. 1 [8], where the distortion caused by interference increases with the spatial density of interferers. The field trial in which the samples were collected included 200 (moving) vehicles equipped with wireless modems, where the test run first ran with 100 active transmitters; then another 50 were added, and in the final third of the test all 200 modems were transmitting 3 parts delineated in Fig 1]. Note that these and other RSSI measurements featured here are made on OFDM transmissions with a 10MHz bandwidth centered near 5.9 GHz, in compliance with V2V DSRC IEEE802.11p, using Atheros 802.11p chips. It appears that fading is increased as more Tx-s are activated with V2V DSRC IEEE802.11p, using Atheros 802.11p chips. The Rx power can also go below the noise floor at any distance as a result of deep fades, due to multi-path. Per Fig 1 the interference causes similar fading on RSSI samples, possibly more intense, causing more losses.

The stochastic EM algorithm is a known approach for computing ML estimates in the mixture problem. Our model is derived from an extension of the SEM algorithm [9], dubbed SEMcm, in a particular case of incomplete data [6], where the information loss is due to both mixture of distributions and censored observations. We aim to estimate the parameters of a left-censored dual mixture, which we propose as a model of observed wireless RSSI samples with countable losses, following [6].

III. BASIC ALGORITHMIC ELEMENTS

A mixture of 2 distributions of the same family \( p(y|\phi_i), i = 1, 2 \), is defined by

\[
p_{\varphi}(y) = \alpha_1 p(y|\phi_1) + \alpha_2 p(y|\phi_2).
\]

Here, \( y \) is the RV modeling an arbitrary mixture sample. \( \alpha_1 \) is the mixing probability. Equivalently,

\[
p_{\varphi}(y, z) = p(y|\varphi z) = f_{\varphi z}(y)
\]

is the joint distribution of the RVs \( Y \) and \( Z \), where \( Z \) is the indicator RV modeling the association with one of the two mixture components (with probability \( \alpha_i \)), and the subscript represents the PDF parameters that we aim to estimate:

\[
\varphi = (\alpha_1, \alpha_2, \phi_1, \phi_2), \ 0 \leq \alpha_1 \leq 1, \ \alpha_1 = 1 - \alpha_2.
\]

We propose that de-logged RSSI values be modeled by a dual mixture of GPDFs \( p(y|\phi_i), i = 1, 2 \). Hence, we have

\[
\phi_i = (m_i, \Omega_i): p(y|\phi_i) = \frac{1}{\Gamma(m_i)\Omega_i} \left( \frac{y}{\Omega_i} \right)^{m_i-1} e^{-\frac{y}{\Omega_i}}.
\]

This model is also depicted in plate notation in Fig. 2.(a).

Next, we introduce censoring: let \( y \in R \) where \( R \) is partitioned into disjoint domains \( R = R_0 \cup R_1 \), where \( R_0 \) is the subset of uncensored data, while \( R_1 \) is the subset corresponding to left-censored data, i.e., \( y \leq c_L \) where \( c_L \) denotes left threshold. Let us assume that there are \( n \) samples total (e.g., transmitted packets), \( r_o \) of which are uncensored (received packets): \( y_k = x_k \in R_0, k \in C_o, |C_o| = r_o \), and \( r_1 \) left censored (lost) samples: \( y_k \in R_1, k \in C_1, |C_1| = r_1 \), where \( r_o + r_1 = n \). Note that \( C_o \) and \( C_1 \) are disjoint sets of sample indices (e.g., packet sequence numbers SNs), \( x_k \) is measured while \( y_k \) is te real value (which are not equal for censored samples). In our model, total number of samples and losses could be obtained by tracking SNs of received packets. We define

\[
T^{(p+1)}_{o,i}(x_k) = E[Z_{k,i}|y = x_k, \phi^{(p)}] = \frac{\alpha_i^{(p)} p(x_k|\phi_i^{(p)})}{\sum_{t=1}^{2} \alpha_t^{(p)} p(x_k|\phi_t^{(p)})},
\]

where \( i = 1, 2, k \in C_o \). \( T^{(p+1)}_{o,i}(x_k) \) denoting current estimate of the probability that uncensored sample \( x_k \) belong
Replacing each missing data by a value drawn at random from $h(y, z_i, \varphi^{(p)})$ (S-step), and then computes the ML estimate based on the completed sample (M-step). We first define the three steps for the left-censored dual-mixture in general, and then present the specific expressions for GPDF.

**E-step:** Compute $T_{o_k}^{(p+1)}(x_k)$ for $k \in C_o$, $i = 1, 2$

Compute $T_{i}^{(p+1)}$ for $i = 1, 2$

**S-step:** (1) For $x_k \in R_o, k \in C_o$ simulate $r_o$ binary

vectors $\tilde{z}_k^{(p+1)} = \{z_k^{(p+1)}, z_k^{(p+1)}\}$ by running Bernoulli experiments w.p. $T_{o_k}^{(p+1)}$; (2) simulate $r_1$ binary vectors $\tilde{z}_L^{(p+1)} = \{z_L^{(p+1)}, z_L^{(p+1)}\}$, $i = 1, \ldots, r_1$, each as a Bernoulli experiment w.p. $T_{i}^{(p+1)}$; (3) simulate $r_l$ missing left censored values sampling from $h(\cdot|c_L, \varphi^{(p)}) = \frac{p_L^{(p)}(\cdot)}{f_L^{(p)}(y)|E_y, \varphi^{(p)}}$;

**M-step:** Choose $\varphi^{(p+1)} = \arg\max_{\varphi \in \Phi} Q(\varphi, \varphi^{(p)})$

where

\[ Q(\varphi, \varphi^{(p)}) = \sum_{i=1}^{2} \left( \sum_{k \in C_a} z_k^{(p+1)} + \sum_{j=1}^{r_1} z_L^{(p+1)} \right) \log \alpha_{i}^{(p+1)} + \sum_{i=1}^{2} \left( \sum_{k \in C_o} z_k^{(p+1)} \log p(x_k|\varphi_{i}^{(p)}) + \sum_{j=1}^{r_l} z_L^{(p+1)} \log p(y_L|\varphi_{i}^{(p)}) \right) \]

We next evaluate $Q(\varphi, \varphi^{(p)})$ for GPDFs, resulting in the proposed channel estimation algorithm, dubbed SEMcmG:

**E-step:** as in (8)-E, based on (4)

**S-step:** as in (8)-S, do (1)-(3), based on (4)

**M-step:** Based on (4) and (9) solve

i. \[ \frac{\partial Q(\varphi, \varphi^{(p)})}{\partial \alpha_{i}} = 0 \Rightarrow \alpha_{i}^{(p+1)} = \frac{1}{n} \left( \sum_{k \in C_o} z_k^{(p+1)} + \sum_{j=1}^{r_l} z_L^{(p+1)} \right) \]

ii. \[ \frac{\partial Q(\varphi, \varphi^{(p)})}{\partial \Omega_{i}^{(p+1)}} = 0 \Rightarrow \Omega_{i}^{(p+1)} = \frac{\Omega_{i}^{(p+1)}}{m_i} \]

\[ \Omega_{i}^{(p+1)} = \frac{\sum_{k \in C_a} z_k^{(p+1)} x_k + \sum_{j=1}^{r_1} z_L^{(p+1)} y_L}{m_i} \]

iii. \[ \frac{\partial Q(\varphi, \varphi^{(p)})}{\partial m} = 0 \Rightarrow \Psi(x) = \frac{\Gamma(x)}{\Gamma(x)} \]

\[ \Psi(x) \approx \log x - \frac{1}{2} - \frac{1}{12x} - \frac{1}{12x^2} - \frac{1}{12x^3} \]

\[ L_{i}^{(p+1)} = \sum_{k \in C_a} z_k^{(p+1)} L_{i,k}^{(p+1)} + \sum_{j=1}^{r_l} z_L^{(p+1)} L_{i,s}^{(p+1)} \]

Solve $\Psi(m_i) = \frac{\Gamma(x)}{\Gamma(x)}$.

Note that we are frequently averaging over the expected number of samples. Total number of samples and losses could be obtained by tracking sequence numbers of received packets.
Fig. 3. The making of left-censored 2-mixture representing a mobile Rx signals (PL) with interference and attenuation.

V. EVALUATION

A. Model Evaluation on Simulated Data

Besides evaluating SEMcmG algorithm on some trivial data sets (one component with left-censoring [11]; one doubly-censored component), we successfully evaluated SEMcmG on a simulated mixture of two left-censored components, which was meant to emulate interference affected RSSI samples. The first component represents the signal over a distance range identical to the range considered in the empirical data evaluation: \( l_d = 23 - 32 \), where \( l_d \) is the log-distance, defined as \( 10 \log_{10}(\text{distance in m}) \). The second component models RSSI samples with strong interference over the same distance range. We simulated different parameters, mostly with the interference component having \( m=1 \) (i.e., \( m_2 = 1 \)), following our discussion in Section I. The results are encouraging. However, we now present a mixture with arbitrary parameters, chosen to create a signal cloud visually distinguishable from the interference cloud in the mixture scatter-plot (bottom left pane in Fig. 3), while capable of exemplifying main concerns about censored RSSI mixtures. The \( m_1 \) is chosen slightly high for the assumed mobile signal (\( m_1 = 7 \)), while \( m_2 = 35 \); such a high value of \( m_2 \) may represent a single (or dominant) interferer.

Signal attenuation over space is exponential, with the attenuation coefficient to be determined through parameter estimation. We choose to present the exponential attenuation in dBm domain as a linear function of \( l_d \). Hence, as in our prior work [8], median path-loss [PL] is fitted by the straight-line function

\[
[PL] = A - Bl_d. \tag{11}
\]

Note that PL is defined as \( PL = RSSI - 10 \log_{10}(P_t) \), where \( P_t \) is the Tx power. Hence, it is distributed as log-Gamma. We present data points in some of our plots as PL rather than RSSI, as it reflects the propagation medium only (independent of Tx power). The simulated \( \Omega \) was chosen so that the linear fit into the dBm value of the Gamma mean (i.e., \( 10 \log_{10}(\text{mean}) \)) vs. Tx-Rx distance be equal to (11) with \( A=-16, B=3 \). Note that \( \Omega \) is a function \( \Omega(l_d) \). With these values, the signal only scatter plot (in dB) is presented in the upper-left plot in Fig. 3. As for interference, for simplicity and without loss of generality, we propose that the median interference is constant over space, e.g., assuming one distant interferer. Such interference points (dBm) are shown in the upper-right plot in Fig. 3.

Notice that for both components we generated points for discrete values of \( l_d \), referred to as distance bins, with 0.5 dB space in between. For each bin we generated 1000 signal (or interference) points, referred to bin arrays. Then, for each bin, and each bins sample index (1-1000) we would select with probability either signal or interference point in that place, making a balanced mixture of the two components, and ending up with 1000 points per bin (bottom-left in Fig. 3). The choice of the mixing coefficient that gives equal weight to both components is deliberate, as such mixtures were hardest to separate. Finally, we censor (drop) the points that are below the threshold \( cL = -109 \) (indicated in bottom plots of Fig. 3), resulting in a set of points in the bottom-right plot of Fig. 3. These are the points fed into SEMcmG, along with the initial values of the parameters, and the information of how many samples per bin were censored. The initial values were distorted with respect to the real values up to 50% in the bottom-right plot of Fig. 3. The red line that was obtained as a Least Square Error (LSE) estimate of the mean of the censored mixture, as opposed to the black line that represents the real mean of the signal. This illustrates how much the assumption of one component (as in the presented LSE) can cost in terms of estimation error. With SEM, the estimates (per bin) were perfect for most simulated mixtures if the data losses constituted less than 60-70% of data, while for higher losses they were just better than ML estimates. For this particular mixture, losses were up to 45% (Fig. 5), in order to highlight
the cluster overlap problem, i.e. the distance bins where the median values of the components were indistinguishable. Please observe the green line in Fig. 4 which illustrates the signals mean estimate. Note that only in the area around \( l_d = 27 \) (cluster overlap) does SEM diverge from the real mean, while following the LSE mean estimate, and in the same area the interference mean estimate follows that of the real signal. We are looking into additional mechanisms to address this phenomenon.

Fig. 5 shows the \( m \) estimate per bin. Again, as the likelihood equations are intractable for any maximum likelihood estimate, we compare our results for the \( m \) parameter with good existing approximations. The \( ML \) and moment-based (MB) estimates in Fig. 5 are calculated based on the \( r \) received samples. The former one is obtained according to the following maximum likelihood approximation

\[
m_{\text{ML}} = 6 + \sqrt{36 + 48\Delta},
\]

where

\[
\Delta = \ln \left( \sum_{i=1}^{r} p_i \right) - \frac{1}{r} \sum_{i=1}^{r} \ln p_i
\]

and \( p_i \) is the Rx power sample (de-logged RSSI). The latter, \( m_{\text{MB}} \), follows eqn. (10) from [12], which is based on the first two sample moments of the received power \( p_i \). The ML and MB estimates never outperform the SEM estimate, even not in the cluster overlap area (Fig. 5). In fact, outside the overlap, ML and MB are producing huge errors, as they assume one Gamma distribution, and, hence, they are interpreting the wide clouds outside of central area as a sign of deep fades; consequently, \( m \) is estimated too low (around 1). This is a very important argument for the proposed approach, as interference clearly cannot be accounted for by any single-component model. A feature of interest for on-line estimation is the convergence rate. We illustrate it in Fig. 6 for both \( \Omega \) and \( m \) and a given bin: a step-by-step evolution of the estimated parameter. It seems that both estimates could have been better if we ran some additional iterations.

Fig. 6. Parameter estimates convergence over 30 iterations to known real values (bin 10): \( \Omega \) - upper plot, \( m \) estimate - bottom.

B. Model Evaluation on Empirical Data

The ZR trial, described in [8], included only one Tx at a time, mounted on a vehicle that traveled back and forth from the static Rx-s on a straight road \( d_{\text{max}} = 1200 \) m long. This scenario with no interference helped us to study the performance of our SEMcm algorithms in terms of signal component estimates, when the initial values for the (non-existent) interference were arbitrary. As the Tx was mobile, suggesting Gamma distribution, SEMcm with Gaussian PDFs gave bad estimates (as expected) and numerical instabilities. SEMcmG showed good results. The initial values for the signal parameters were taken from imperfect estimates, based on the linear LSE fit into a pathloss function that was linear only beyond a break point, and also due to noise-floor saturation (Fig. 7 upper pane).

For simplicity we performed SEMcmG only for the distance bins after the break-point (2nd segment), as the smaller distances involved the two-ray phenomenon. The linear fit of the initial \( \Omega m \) in dBm, represented by the yellow line in Fig. 7 matches (11) with coefficients A2 and B2. Other coefficients, based on the LSE over the 2nd segment only, came closer to the real median PL (known from running the same field trial with higher Tx power, which avoids the noise floor within traversed distances).

The SEMcm estimated line (red line with circular markers) is almost the same as the real one. The initial value for \( m \) was 1.5 (bottom black line in the middle pane of Fig. 7), yet
SEMcmG managed to improve it to 3.3 on average, which is identical to its ML estimate. Now, the ML estimate works optimally when there is sufficient number of samples, which was the case here. The bottom pane of Fig. 7 shows the number of transmitted packets in black, and the number of lost packets in red. The last bin has the worst losses (75%), yet, more than 500 packets received is sufficient for ML.

In conclusion, without interference, SEMcm outperforms the LSE approach in estimating the mean (10log(Ω(m))), while it is comparable to ML in estimating m. Finally, we present Fig. 8 which is based on the data featured in Fig 1. Apart from show-casing the notion of dual mixture and censoring, this figure affirms the censored mixture approach, as it illustrates a good match between the SEM-reconstructed PDF of the data featured in Fig 1, and its empirical distribution. Observe that the points left of the black vertical line around -115dBm represent censored samples (i.e, $c_L = -115$).

VI. CONCLUSION

Our main contribution is a novel model of interference affected RSSI samples, presented as censored mixture of Gamma PDFs, based on the insight from data collected for varying interference levels (see Fig. 1). Also, we applied a flavor of EM algorithm which not only mechanizes the computation of the parameters’ ML estimates for our complex statistical model of incomplete non-Gaussian mixed data [13], [14], but also utilizes stochastic randomization to avoid strong dependence on its starting position, convergence to a saddle point, and low convergence rate. A great property of this method is that it leverages the count of lost data, to improve estimates for small number of samples, which is especially important for online estimation based on crowd-sourced data.

Our future work will explore online versions of EM algorithms [15] applied to our problem. Also, future work will address improvements for signal levels that are on average too close to interference levels, such as in cluster-overlap bins in Figures 4 and 5. Although this is a common problem in data clustering, we believe that good predictive models for cluster overlaps could be developed based on signal samples in distance bins with good separation.

REFERENCES