Clustering – Gaussian mixture model (2)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

March 6, 2024

Agenda of today's class

- Discussed Gaussian mixture modelling and started developing EM algorithm
- Finish developing EM algorithm
  - Reference: Chapter 6 of FCML

Mixture model – generative process (recap)

- We assume that data is generated by multiple Gaussians we propose
- Two-step procedure for sampling the nth data object $\mathbf{x}_n$:
  - 1) Select one of the three Gaussians; 2) Sample $\mathbf{x}_n$ from this Gaussian
- Step 1 chooses one value from a discrete set, like rolling a die
  - To do this, we just need to define the probability of each outcome π_k such that $\sum_k \pi_k = 1$
- We used z_{nk} as an indicator variable
  - If we choose kth component as the source of nth object, we set $z_{nk} = 1$, and $z_{nj} = 0$ for all $j \neq k$
- Require likelihood of data objects $\mathbf{x}_n$ under the whole model:$p\left(\mathbf{x}_n \mid \Delta, \boldsymbol{\pi}\right)$
- We started with likelihood of a particular data object conditioned on $z_{nk} = 1$:

$$p\left(\mathbf{x}_n \mid z_{nk} = 1, \Delta\right) = p\left(\mathbf{x}_n \mid \Delta_k\right)$$

- Summed both sides over k (marginalising over the individual components) yields

$$p\left(\mathbf{x}_n \mid \Delta, \boldsymbol{\pi}\right) \;=\; \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \Delta_k\right)$$

- Made standard independence assumption and extended this to likelihood of all N data objects:

$$p(\mathbf{X} \mid \Delta, \boldsymbol{\pi}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \Delta_k\right)$$

Simplification of log likelihood using Jensen inequality (recap)

- We want to maximise the log likelihood

$$L = \log p(\mathbf{X} \mid \Delta, \boldsymbol{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right))$$

- Summation inside logarithm makes finding optimal parameter $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}$ difficult
- EM algorithm overcomes this problem by deriving a lower bound on this likelihood
  - Instead of maximising L directly, we maximise a lower bound obtained using Jensen inequality
- To obtain lower bound, we multiply and divide expression inside summation over k by q_{nk}
  - $q_{nk} \geq 0$ with $\sum_{k=1}^{K} q_{nk} = 1$
  - q_{nk} can be considered as some probability distribution over K components for the nth object
- We used the Jensen inequality and got a following lower bound

$$L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \frac{q_{nk}}{q_{nk}} \geq \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \left(\frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right) = \mathcal{B}$$

- Maximize lower bound $\mathcal{B}$ to calculate $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}, q_{nk}$
  - Partially differentiate $\mathcal{B}$ w.r.t $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}, q_{nk}$ and set it to zero

Updating π_k (1)

- Bound
$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$
- We first update π_k

$$\begin{aligned}
\mathcal{B} &= \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) + \cdots \\
\frac{\partial\mathcal{B}}{\partial\pi_k} &= \frac{\sum_{n=1}^{N} q_{nk}}{\pi_k} - \lambda = 0 \Rightarrow \sum_{n=1}^{N} q_{nk} = \lambda\pi_k \qquad (1)\\
\sum_{k=1}^{K}\sum_{n=1}^{N} q_{nk} &= \lambda\sum_{k=1}^{K}\pi_k \overset{(a)}{\Rightarrow} \sum_{n=1}^{N} 1 = \lambda \Rightarrow \lambda = N
\end{aligned}$$

- Arrow (a) used the fact that $\sum_{k=1}^{K} q_{nk} = 1$ and $\sum_{k=1}^{K}\pi_k = 1$ by definition.
- Substituting $\lambda = N$ into Eq. (1) gives us the expression for $\pi_k = \frac{1}{N}\sum_{n=1}^{N} q_{nk}$

Updating $\boldsymbol{\mu}_k$ (1)

- Bound

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$

- Updating $\boldsymbol{\mu}_k$: only second term of $\mathcal{B}$ includes $\boldsymbol{\mu}_k$ – expand multi-variate Gaussian $p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$

$$\begin{aligned}\mathcal{B} &\propto \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\left(\frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)\right)\right)\\ &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\left((2\pi)^d\left|\boldsymbol{\Sigma}_k\right|\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)\end{aligned}$$

- Making use of the identity $\left(f(\mathbf{w}) = \mathbf{w}^T\mathbf{C}\mathbf{w}, \quad \frac{\partial f(\mathbf{w})}{\partial\mathbf{w}} = 2\mathbf{C}\mathbf{w}\right)$

$$\frac{\partial\mathcal{B}}{\partial\boldsymbol{\mu}_k} = -\frac{1}{2}\sum_{n=1}^{N} q_{nk}\times\frac{\partial\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)}{\partial\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)}\times\frac{\partial\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)}{\partial\boldsymbol{\mu}_k} = \sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)$$

Updating $\boldsymbol{\mu}_k$ (2)

- Equating to zero and rearranging gives us an expression for $\boldsymbol{\mu}_k$

$$
\begin{aligned}
\sum_{n=1}^{N} q_{nk} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) &= 0 \\
\sum_{n=1}^{N} q_{nk} \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_n &= \sum_{n=1}^{N} q_{nk} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\
\sum_{n=1}^{N} q_{nk} \mathbf{x}_n &= \boldsymbol{\mu}_k \sum_{n=1}^{N} q_{nk} \\
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} q_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} q_{nk}}
\end{aligned}
$$

Updating $\boldsymbol{\Sigma}_k$ (1)

- As with $\boldsymbol{\mu}_k$, we only need to look at the multi-variate Gaussian $p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ term of $\mathcal{B}$

$$\mathcal{B} \propto -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \log\left((2\pi)^d |\boldsymbol{\Sigma}_k|\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)$$

- Ignoring the constant (2π) part of the first term, we are left with

$$\mathcal{B} \propto -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \log\left(|\boldsymbol{\Sigma}_k|\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)$$

- To take partial derivatives with respect to the matrix $\boldsymbol{\Sigma}_k$, we need two identities

$$\frac{\partial \log|\mathbf{C}|}{\partial \mathbf{C}} = (\mathbf{C}^T)^{-1} \text{ and } \frac{\partial \mathbf{a}^T\mathbf{C}^{-1}\mathbf{b}}{\partial \mathbf{C}} = -(\mathbf{C}^T)^{-1}\mathbf{a}\mathbf{b}^T(\mathbf{C}^T)^{-1}$$

- We take partial derivatives with respect to $\boldsymbol{\Sigma}_k$

$$\frac{\partial \mathcal{B}}{\partial \boldsymbol{\Sigma}_k} = -\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} + \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}$$

- Above equation uses the fact that $\boldsymbol{\Sigma}_k$ is symmetric and therefore $(\boldsymbol{\Sigma}_k)^T = \boldsymbol{\Sigma}_k$

Updating $\boldsymbol{\Sigma}_k$ (2)

- We now equate $\frac{\partial \mathcal{B}}{\partial \boldsymbol{\Sigma}_k}$ to zero

$$-\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} + \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} = 0$$

$$\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} = \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}$$

- Pre- and post-multiplying both sides by $\boldsymbol{\Sigma}_k$ allows us to cancel all of the $\boldsymbol{\Sigma}_k^{-1}$:

$$\begin{aligned}
\boldsymbol{\Sigma}_k \sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k &= \boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^{-1}\sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k \\
\boldsymbol{\Sigma}_k \sum_{n=1}^{N} q_{nk} &= \sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} q_{nk}}
\end{aligned}$$

Updating q_{nk} (1)

- Bound
$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$
- Updating q_{nk}, which appears all three terms in $\mathcal{B}$. It is subject to the constraint $\sum_{k=1}^{K} q_{nk} = 1$
- Using Lagrangian method we have
$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_K\right) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk} - \lambda\left(\sum_{k=1}^{K} q_{nk} - 1\right)$$
- Taking partial derivatives with respect to q_{nk} gives
$$\frac{\partial\mathcal{B}}{\partial q_{nk}} = \log\pi_k + \log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) - \left(1 + \log q_{nk}\right) - \lambda$$
- Setting to zero, rearranging and exponentiating gives us an expression for q_{nk}:
$$\begin{aligned} 1 + \log q_{nk} + \lambda &= \log\pi_k + \log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\ \exp\left(\log q_{nk} + (\lambda + 1)\right) &= \exp\left(\log\pi_k + \log p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right) \\ q_{nk}\exp(\lambda + 1) &= \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \end{aligned} \tag{2}$$

Updating q_{nk} (2)

- We need to find the constant term $\exp(\lambda+1)$, we sum both sides over k:

$$\begin{aligned}\exp(\lambda+1)\sum_{k=1}^{K} q_{nk} &= \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\\ \exp(\lambda+1) &= \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\end{aligned}$$

- Substituting above in Eq. (2), we have

$$q_{nk} = \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

Summary and intuitions from the derived expressions (1)

- Four update equations are

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} q_{nk}, \qquad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} q_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} q_{nk}}, \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} q_{nk}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T}{\sum_{n=1}^{N} q_{nk}}$$

$$q_{nk} = \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

- First three expressions rely heavily on q_{nk} - what does q_{nk} represent?
  - Could be interpreted posterior probability of object n belonging to class k

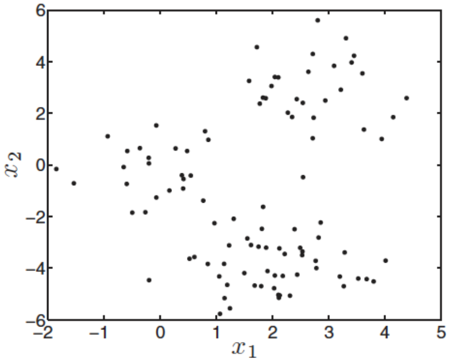
$$p(z_{nk} = 1 \mid \mathbf{x}_n) = \frac{p\left(z_{nk} = 1\right) p\left(\mathbf{x}_n \mid z_{nk} = 1\right)}{\sum_{j=1}^{K} p\left(z_{nj} = 1\right) p\left(\mathbf{x}_n \mid z_{nj} = 1\right)} = \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} = q_{nk}$$

- π_k is the average of all posterior probabilities of belonging to class k
  - Equivalently expected proportion of data belonging to class k
- $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are mean and variance of each class
  - Calculated by weighting each object by its posterior probability of belonging to class k

Intuition from the derived expressions (2)

- Keeping the previous discussion in mind, we can split the four updates into two steps
  - Step-1: update current estimates of model parameter π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ by fixing q_{nk}
  - Step-2: update assignments q_{nk} to reflect the new values of π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$
- Algorithm is very similar to K-means algorithm
  - Updating q_{nk} is analogous to updating z_{nk} in K-means
  - Updating $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ $\boldsymbol{\pi}$ is analogous to updating $\boldsymbol{\mu}_k$ in K-means
- Two key differences from K-means
  - Compute posterior probabilities of cluster memberships rather than making hard assignments and
  - Inferring the component covariances
- Four update equations make up an example of the EM algorithm
- First three updates π_k $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ make up the so-called 'M' (maximisation) step
  - Step maximized the bound conditioned on the values of q_{nk}
- Update of q_{nk} is known as the 'E' (expectation) step
  - Computes expected value of unknown assignments, z_{nk} – we have not derived them this way

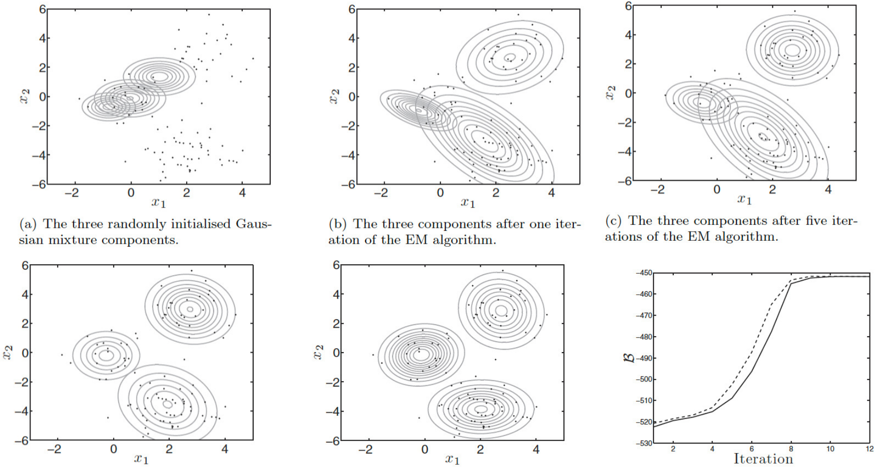
Example of EM algorithm (1)



- Before start performing updates in equations we need to initialise some of the parameters
  - We set $K = 3$ and randomly choose the means and covariances of $K = 3$ mixture components
  - To compute q_{nk}, we need to initialize π_k – uniform distribution over three components: $\pi_k = 1/K$

Example of EM algorithm (2)

- Three resulting Gaussian pdfs are



(a) The three randomly initialised Gaussian mixture components.

(b) The three components after one iteration of the EM algorithm.

(c) The three components after five iterations of the EM algorithm.

(d) The three components after seven iterations of the EM algorithm.

(e) The three components at convergence of the EM algorithm.

(f) The evolution of the bound $\mathcal{B}$ (solid line, Equation 6.8) and log-likelihood L (dashed line, Equation 6.5).

Example of EM algorithm (2)

- We are not interested in Gaussians themselves but assignments of objects to cluster components
  - Provided by values of q_{nk} posterior probability of objects belonging to components
- Consider an object n that has the following values of q_{nk} at convergence:

$$q_{n1} = 0.53, q_{n2} = 0.45, q_{n3} = 0.02$$

- If we must assign it to a particular cluster, first one is most appropriate
  - We are throwing away useful information about relationship object n has with component 2
- Clusterings produced by K-means and mixture model are similar and K-means can be kernelised
  - Mixture models have advantages over K-means due, predominantly, to their probabilistic nature

Clustering in 5G wireless systems - examples

8422 IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. 67, NO. 12, DECEMBER 2019

An EM-Based User Clustering Method in Non-Orthogonal Multiple Access

Jie Ren, *Student Member, IEEE*, Zulin Wang, *Member, IEEE*, Mai Xu, *Senior Member, IEEE*, Fang Fang, *Member, IEEE*, and Zhiguo Ding, *Senior Member, IEEE*