B2.3—*Gamma Expectation Maximisation* (*GEM*) (*v1.1*)

Aim: Cluster and (γ) mixture modelling for IED data

Input(s): IED distributions (aggregated)

Output: γ(α,β) and weight parameter estimates, log-likelihood trace

Req(s): MATLAB (2017a)

B2.3.1—(γ) Parameter Estimation (MLE)

GEM constitutes a (γ) distribution specific application of the expectation maximisation (EM) algorithm. EM is an iterative method to obtain maximum likelihood estimates (MLE) for statistical, mixed models that contain latent values (e.g. class II CO frequency) (Do & Batzoglou 2008). *GEM* is designed as a callable function and a standalone package:

GEM(*IEDdata*,$n_{COMPONENTS}$,*maxIter*,*error_thresh*,*init_mode*,*init_alpha*,*init_beta*,*init_weight*)

MLE derivation of mixed γ(α,β) parameters is well established (Webb 2000; Destrempes et al. 2011). The likelihood function L(X|α|β) is key to statistical inference, describing the collective likelihood that the observed data (X_I) arose from the probability density function f(α,β) (EQN 1.2) of the proposed or fitted quantitative model. Maximum likelihood estimation seeks to maximise likelihood by obtaining γ(α,β) values that best fit (X_I). A more mathematically convenient form is the log likelihood function, the natural logarithmic transformation of EQN 1.2 (EQN 1.3).

$$L(X|\alpha,\beta) = \prod_{i=1}^{N} f(x_i;\alpha,\beta) \quad \textbf{(EQN 1.2)}$$

For N independently, identically distributed (i.i.d) variables (x1….., xN)

$$logL(x_i|\alpha_j,\beta_j) =$$

$$(\alpha_j - 1)\sum_{i=1}^{N} log(x_i) - \sum_{i=1}^{N}\frac{x_i}{\beta_j} - N\alpha_j \log(\beta_j) - Nlog(\Gamma(\alpha_j)) \quad \textbf{(EQN 1.3)}$$

By obtaining the derivative, setting the equation to equal zero and finding the maximum with respect to γ(β), it can be shown that γ(β) estimation can be fully expressed in terms of (X_I) and γ(α) (EQN 1.4). Substituting EQN1.4 into EQN1.2 and subsequently taking the derivative, setting the

equation to equal zero and finding the maximum with respect to γ(α), the equation for γ(α) MLE is obtained (EQN 1.5, 1.6)—where ψ equals the digamma function (EQN 1.7). No closed form solution for γ(α) exists, however, f(x) = log(x) -ψ(x) is numerically well behaved and therefore γ(α) can be estimated through numerical means. MLE γ(β) values are subsequently obtained using the obtained maximised γ(α) value.

$$\beta_j = \frac{1}{\alpha_j} \frac{\sum_{i=1}^{N} \gamma_{i,j} x_i}{\sum_{i=1}^{N} \gamma_{i,j}}$$ **(EQN 1.4)**

Where γ(i,j) denotes the probability of X(i) (data) belonging to cluster j

$$log(\alpha) + \psi(\alpha) = log\left(\frac{\Sigma_i^N \gamma_{i,j} x_i}{\Sigma_i^N \gamma_{i,j}}\right) - \left(\frac{\Sigma_i^N \gamma_{i,j} log x_i}{\Sigma_i^N \gamma_{i,j}}\right)$$ **(EQN 1.5)**

$$log\left(\frac{\Sigma_i^N \gamma_{i,j} x_i}{\Sigma_i^N \gamma_{i,j}}\right) - \left(\frac{\Sigma_i^N \gamma_{i,j} log x_i}{\Sigma_i^N \gamma_{i,j}}\right) - log(\alpha_j) + \psi(\alpha_j) = 0$$ **(EQN 1.6)**

$$\psi(x) = \frac{d}{dx} ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$ **(EQN 1.7)**

B2.3.2—Cluster Analysis

GEM initially segregates data into ($n_{COMPONENTS}$) number of soft clusters (e.g. 2) and subsequently utilises MLE to reiteratively improve the fit of each sub distribution and the overall model, recalculating the log likelihood in a cyclical fashion until a termination criteria is met such as an error threshold or maximum allowed iterations (Figure 2.23A). A useful property of EQN1.6 is that the solution adopts a (-) value if the γ(α) estimate is below the maximised value, and correspondingly a (+) value if above (Figure 2.23B). Via MATLAB function *fzero*, which attempts to find a point (x) where fun(x) = 0 based on sign change, EQN 1.6 is numerically evaluated over a given range of γ(α) values for each subpopulation—a range periodically shifted based on the evaluative outcome. Such a process allows *GEM* to narrow in on the best fit γ(α) value. The relative contribution of each sub distribution (i.e. weight), is estimated by approximating the number of data points which are likely belong to each cluster.

B2.3.3—Parameter Initiation

GEM provides two parameter initiation methods: (i) A kmeans++ algorithm—an established method of parameter initiation (Blömer & Bujna 2013). Kmeans++ initially assigns a centre point to a given number of clusters (nC), assigns data an identity denoting which cluster it belongs to and directly approximates $\gamma(\alpha,\beta)$ parameters of each cluster via method of moments (EQN 1.8) (ii) A biased, non-automated approach whereby the user specifies initial $\gamma(\alpha,\beta)$ and/or weight values for a given number of clusters (nC). The choice of initiation method is context dependent, as shown in (Section 2.20).

$$\alpha = \left(\frac{x}{s}\right)^2 \qquad \beta = \frac{s^2}{x}$$

Where S = Standard Deviation, x = Sample Mean

(EQN 1.8)