README file for "Model-based clustering based on sparse finite Gaussian mixtures", by Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün

The folder 'Code_SpMix' contains 4 files with codes written in R (version 3.0.0) to estimate a sparse finite Gaussian mixture model as defined in Sections 2,3, and 4 of the paper. The code was tested on the Windows 7 platform. The following R packages are required: **bayesm**, **Runuran**, **MASS**, **MCMCpack**, **mclust**, **e1071**, **mvtnorm**, **flexclust**.

The main file is `Analysis_SpMix.R`. It consists of the following procedure steps:

1. Data is generated according to simulation setup I (Section 5.1). Alternatively the 'Crabs data set' can be read.

2. Simulation and prior parameters are specified.

3. The function $MultVar\_NormMixt\_Gibbs\_IndPriorNormalgamma()$ calls the Gibbs sampling procedure to estimate the posterior distribution of the sparse normal mixture model. If $priorOnE0 = TRUE$, the hyperprior $e_0 \sim \mathcal{G}(a, b)$ is specified, otherwise e_0 is set to the specified initial value. If $lambda = TRUE$, the hyperprior $\lambda \sim \mathcal{G}(\nu, \nu)$ is specified, otherwise $lambda = 1$ is fixed.
   The returned value of the function consists of:

   - Mu: draws of the component means,
   - $Sigma$: draws of the component covariance matrices,
   - Eta: draws of the mixture weights,

- S_alt_matrix: matrix consisting of the component allocations S of the observations,
- B: matrix with the shrinkage factors lambda,
- $e0\_vector$: e_0 draws,
- acc_rate: acceptance rate of the MH step for estimating e_0,
- Nk_matrix_alt: number of observations allocated to the components,
- $nonnormpost\_mode\_list$: list of the posterior mode values.

4. After some diagnostic plots the mixture model is identified. First the number of non-empty components $K0$ is determined. Then the draws corresponding to $K0$ are selected and clustered in the point process representation (function $MultVar\_clust\_FS\_FAST\_2()$).

5. Finally, the classification of the observations is estimated and evaluated using error rate, adjusted Rand index and scatter plots.