The function 'plate_model_loglikelihood_evaluation2.R' calculates the log-likelihood scores according to the model described in the article Tiong K.-L. *et al*. "Assessing transcriptomic heterogeneity of single-cell RNASeq data by bulk-level gene expression data". R 4.0.5 version was used for running this code.

The function is supposed to be called from an external R code, for example as following:

```
source("plate_model_loglikelihood_evaluation2.R")
output <-
plate_model_loglikelihood_evaluation2(pmode,Q,P,bulkdata,scdata,sccdfs,genegroups,bulksubtypeinds,cellsubtypeinds,cellcelltypeinds,datamode,bdw)
```

The parameters fed into the function are either real numbers or matrices of real numbers. In the case of matrices, it is necessary to represent them as the 'matrix' class and not as 'data.frame'. For example, if the input data are read from the corresponding csv files, the following type conversion should be used:

```
Q <- as.matrix(read.csv("Q.csv",header=FALSE))
```

The function produces the following outputs: joint log likelihood score of scRNAseq data, the log likelihood scores of each cell and each cell type, the inferred cell types of each cell.

The function takes the following inputs.

pmode

An integer number which allows for four alternative ways to estimate GP, Pr(expression of a gene | gene group, cell type).

pmode=1: Estimate the pdf of a gene group from Q and bulk expression data.
pmode=2: Estimate the pdf of a gene group from the distribution in single-cell RNAseq data.
pmode=3: Estimate the pdf of a gene group from Q and single-cell RNAseq data.
pmode=4: Directly load GP and store it in Q.

Q

An $n \times k$ signature matrix Q denoting the expression profiles of n genes in k cell types (for pmode=1,2,3) or Q=GP (for pmode=4). In the latter case, it is a $c \times k \times l$ tensor for c gene groups, k cell types, and l intervals of gene expression values.

P

An $k \times m$ mixture coefficient matrix P denoting the proportions of k cell types in m bulk samples, where the P entries in each column are nonnegative and sum to 1.

bulkdata

An $n \times m$ matrix of bulk gene expression data (TPM counts) of n genes in m bulk samples.

scdata

An n×j matrix of single-cell RNAseq data (TPM counts) of n genes in j single cells.

sccdfs

An n×j matrix of single-cell RNAseq cdfs, where each row is a cdf transform of the corresponding row in 'scdata'.

genegroups

A 1×n matrix of gene group labels, where n is the total number of genes, and all genes belonging to the group '1' are labeled by integer 1, genes belonging to the group '2' are labeled by integer 2, etc.

bulksubtypeinds

A 1×m matrix of subtype labels of bulk samples, where m is the total number of bulk samples, and all samples belonging to the subtype '1' are labeled by integer 1, samples belonging to the subtype '2' are labeled by integer 2, etc.

cellsubtypeinds

A 1×j matrix of subtype labels of cells, where j is the total number of cells, and all cells belonging to the subtype '1' are labeled by integer 1, cells belonging to the subtype '2' are labeled by integer 2, etc.

cellcelltypeinds

A 1×j matrix of cell type labels of single cells, where j is the total number of cells, and all cells belonging to the type '1' are labeled by integer 1, cells belonging to the type '2' are labeled by integer 2, etc.

datamode

An integer number which allows for two ways to incorporate entries to evaluate log likelihood scores.

datamode=1: Use all entries to evaluate log likelihood scores.
datamode=2: Use only positive entries to evaluate log likelihood scores.

bdw

The value of 'width' parameter of kernel density estimation. If bdw=0, the default value is used.