BFRMFactor

Jeffrey T. Chang, Carlos Carvalho, Joseph E. Lucas, Quanli Wang,
Joseph R. Nevins, and Mike West
3 January 2011

Summary: This module performs a Bayesian latent factor decomposition on a gene expression data set. Briefly, it models a gene expression data set as being comprised of a set of latent factors. The decomposition can be initiated on a subset of the genes (the nucleus), and addition genes and factors are added in a statistically principled *evolutionary search*.

For more information on BFRM, see:
> Carvalho C, *et al*. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*. 103:1438-1456, 2008.

For more information on its application in modeling pathways, see:
> Chang JT, *et al*. A Genomic Strategy to Elucidate Modules of Oncogenic Pathway Signaling Networks. *Molecular Cell*. 34(1): 104-114, 2009.

Parameter: dataset
This is the gene expression data set to be analyzed. This should be in PCL, GCT, or some other data formats.

Parameters: filter_mean, filter_var
This module can filter the data set to remove genes with low expression or genes that do not vary across the data set. *filter_mean* should be a value between 0.0 and 1.0 that indicates the portion of genes to remove before analysis. For example, a value of 0.25 will remove the 25% of genes with lowest mean expression. Similarly, *filter_var* is a value between 0.0 and 1.0 that indicates the portion of genes to remove, based on their variance across the data set.

Parameter: num_control_vars
This module can remove noise from the data set before analysis. In brief, it will perform an SVD on the affymetrix control probe sets (those that start with *affx-*). Assuming that the control probe sets do not vary across the data set, any structure seen in the principal components should be noise. BFRM will *model out* that structure. *num_control_vars* is the number of principal components of noise to remove. The more control variable used, the more noise is removed, at the expense of potentially removing signal.

If this feature is used, the data set must have been generated on Affymetrix microarrays, and the control probe sets must not have been removed from the data set.

Parameters: nucleus_file, start_factors, max_factors, max_genes
These parameters govern the evolutionary search. *nucleus_file* is a text file where each line contains the name (or probe set ID) of a gene to nucleate the evolutionary search. These names (or IDs) must be given in the data set. *start_factors* is the number of

principal components to use in the first iteration of the model. Typically this is set at 1. *max_factors* is the maximum number of factors to include in the search. Once this number of factors is reached, the search is stopped. *max_genes* is the maximum number of genes to include in the search. Once this number of genes is reached, the search is stopped. This controls how far away from the nucleus genes that the model will search. For factors that are more closely related to the nucleus, this is typically set to 500. For deeper searches, this is set to 1000.

For help, please email: jeffrey.chang@duke.edu.