Sparse Factor Analysis for Detecting Copy Number Variations (CNVs)

Andreas Mitterecker, Djork-Arne Clevert, Andreas Mayr, An De Bondt, Willem Talloen, Marianne Tuefferd, Hinrich Göhlmann, Sepp Hochreiter

Recently, CN.FARMS has been proposed for analyzing copy number variations (CNVs) with oligo genotyping arrays like the Affymetrix SNP6 chips. CN.FARMS extracts a common hidden factor of neighbouring reporters on the DNA, where the factor represents the local copy numbers. If an individual has more copies of a DNA-subsequence compared to a reference, then reporters on this subsequence have higher intensity than those of the reference. CN.FARMS assigns to each copy number estimate a confidence value (signal likelihood) which is large if reporters agree to each other across the samples.

Standard factor analysis assumes a Gaussian factor distribution which, however, is a wrong assumption for CNVs. Redon et al. 2006 discovered that most CNVs affect less than three individuals out of the 269 HapMap samples. Thus, CNV analysis requires the assumption of the hidden factor being sparsely distributed.

To account for a sparse hidden variable, we assume either a Laplace or a multimodal distribution. Because the posterior cannot be computed analytically, we use for the Laplace prior both a variational and an approximated EM approach. For the multimodal distribution we use a lower bound on the likelihood and represent the prior by a mixture of Gaussians.

The sparse CN.FARMS methods were applied to Affymetrix SNP6 chips on the HapMap data set. We noticed that many previously reported CNVs are likely to be false positives and also found novel CNVs.