Detecting rare copy number variations (CNVs) with sparse coding

Andreas Mitterecker[1], Djork-Arné Clevert[1,3], Andreas Mayr[1], An De Bondt[2], Willem Talloen[2], Hinrich Göhlmann[2], Sepp Hochreiter[1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, 4040 Linz, Austria
[2] Johnson & Johnson Pharmaceutical Research & Development. A division of Janssen Pharmaceutica, Beerse, Belgium
[3] Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany

Motivation: High-density oligonucleotide genotyping microarrays, especially Affymetrix SNP6 chips, are widely used for high-resolution copy number analysis. The spiraling list of new identified copy number variations (CNVs) which are associated with human disease, evidence suggests that CNVs are extremely relevant in medical research. In order to identify CNVs more reliable, we have proposed a Maximum a posteriori factor analysis model called cn.FARMS. The latent variable, the factor, captures the simultaneous increase or decrease of DNA amount at neighboring chromosome locations measured by the intensity of oligonucleotide probes. This increase or decrease indicates amplification or deletion of a DNA region that is a CNV. cn.FARMS considerably reduces the false discovery rate (FDR) by combining adjacent chromosome locations to an ensemble voting (agreement of multiple measurements) instead of relying on a single measurement as other method do.

Nevertheless, standard cn.FARMS assumes that the latent variable is Gaussian distributed, which implies that the distribution of amplifications and deletions are also Gaussian distributed. However Redon et al. 2006 showed that most CNVs affect less than three individuals out of 270 HapMap samples. These rare events are hard to detect by cn.FARMS as they would be interpreted as noise. An appropriate approach would model those changes by a sparse factor, meaning that the factor takes for most cases its default value (CN 2) and deviates only in few cases considerably from this value. Therefore we propose a factor analysis model with a Laplacian prior, which leads to a sparse factor distribution. But now we face another problem: the likelihood much harder to compute. We tackled this problem by applying an algorithm that employs a variational expectation maximization algorithm to the sparse prior, which optimizes a lower bound on the likelihood and is based on a local Gaussian approximation to the mode of the Laplacian prior distribution. We have also developed an exact approach.

Results: We have applied the Laplacian cn.FARMS model on the HapMap dataset to detect CNVs. We could verify most of published copy number variable regions and found new ones. However some known CNVs seem to be false positives.

The Model

$$\boldsymbol{x} \;=\; \boldsymbol{\Lambda} z \;+\; \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \;\sim\; \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Psi}\right)$$

z : hidden factor due to CNVs

$\boldsymbol{\epsilon}$: independent noise

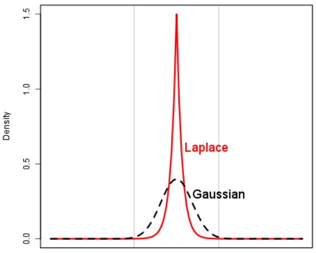
$\boldsymbol{\Psi}$: diagonal covariance matrix (independent noise)

z and ϵ are independent

Reasons for Sparse Factors

- Few components are significantly activated
- Large values are more probable than with Gauss ($p(z) = \frac{1}{2} e^{-|z|}$)
- Most values at the mode

Sparse Factors by prior $p(z)$ as
Laplace distribution ($p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$)



Problem

Likelihood: $p\left(\boldsymbol{x} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}\right) \;=\; \int p\left(\boldsymbol{x} \mid z, \boldsymbol{\Lambda}, \boldsymbol{\Psi}\right)\; p\left(z\right)\; dz$

is non-Gaussian for Laplace prior which makes computation more difficult

Solution 1: Variational approach

Lower bound on the Likelihood by introducing distribution Q(z)

$$\begin{aligned}
\log\; p(\boldsymbol{x}) \geq \log\; p(\boldsymbol{x}|\xi) &= \int Q(z)\; \log\; p(\boldsymbol{x}|\xi) dz \\
&= \int Q(z)\; \log\; \frac{Q(z)}{p(z|,\xi)} dz - \int Q(z)\; \log\; \frac{Q(z)}{p(z,|\xi)} dz \\
&\geq \int Q(z)\; \log\; p(z, \boldsymbol{x} \mid \xi) dz \;=\; B
\end{aligned}$$

B = likelihood: $Q(z) = p(z|\boldsymbol{x}, \boldsymbol{\Lambda}^{new})$

Standard EM: $Q(z) = p(z|\boldsymbol{x}, \boldsymbol{\Lambda}^{old})$ (B = likelihood after M step)

Variational approach: $\arg\max_{\xi}\; p(\boldsymbol{x}|\xi) = \log p(\boldsymbol{x})$

Solution 2: Exact computation

Based on truncated Gaussian moments

Pipeline

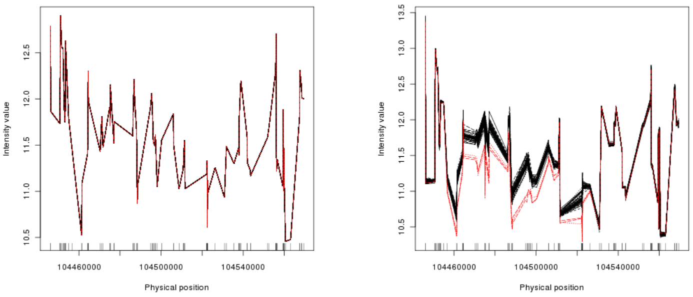
➔ Probe level data
1. Normalization
   - sparse.FARMS
   - ACC
   - Quantile
   - VSN
2. Correction for sequence effects
3. Allele signals correction
PM_A + PM_B
4. Single locus modeling
   - Laplace FARMS
   - Medianpolish
   - MBEI
5. Fragment length correction
6. Multi loci modeling
   - Laplace FARMS
➔ Raw copy number

Experiments on HapMap Dataset

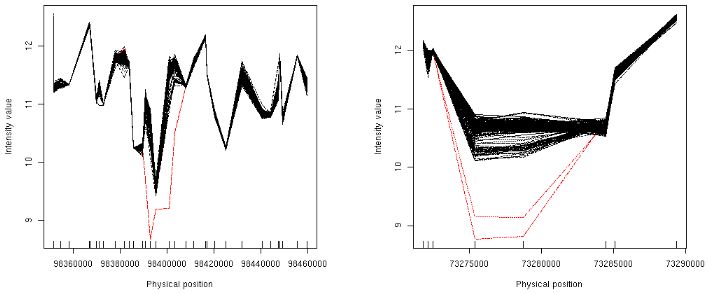
- Dataset: 270 HapMap individuals evaluated by Affymetrix SNP6 microarrays
- Multiloci window mode: combine adjacent probe sets to one measurement
- Reference results: McCarroll et al. 2008

Results and Justification

cn.FARMS with Gaussian prior (left graph) vs. cn.FARMS with Laplace Prior (right graph)
Deletions on three individuals can clearly be detected by Laplacian cn.FARMS
(chr5:104,464,614-104,506,104 – confirmed deletion region)



Novel CNVRs found by cn.FARMS with Laplacian prior


Chr4:98,392,948-98,400,774

Chr6:73,275,400-73,278,772

Conclusion

We have showed that a Laplace prior for cn.FARMS is superior to a Gaussian prior at detecting sparse CNVRs.

Most of the copy number variable regions from McCarroll et al. 2008 could be confirmed by our approach while we also detected novel CNVRs.

From the top ranked 744 regions found by Laplace cn.FARMS but only partly by McCarroll et al. 2008 , 678 are rediscoveries as they are reported in the Database of Genomic Variants --- the remaining are new discoveries.

These results verify our approach to detect rare CNVRs by Laplacian priors.

References

Mark Girolami. A variational method for learning sparse and overcomplete representations. Neural Comput., 13(11):2517-2532, 2001.
Redon et al. Global variation in copy number in the human genome. Nature, 444(7118):444-454,2006.
McCarroll et al. Integrated detection and population-genetic analysis of snps and copy number variation. Nat Genet, 40(10): 1166-1174, 2008.
Zhong et al. An EM algorithm for learning sparse and overcomplete representations. Neurocomputing, 57:469-476, 2004.