

A Unified Method for CNV Detection and Association with Gene Expression

Un Metodo Unificato per l'Identificazione di CNV e l'Associazione fra CNV e l'Espressione Genica

Mahlet G. Tadesse, Alberto Cassese, Michele Guindani, Francesco Falciani and Marina Vannucci

Abstract Copy number variants (CNV) are chromosomal aberrations resulting in DNA segments having an abnormal number of copies. The problem of detecting CNVs has received a lot of attention and several methods have been developed to infer CNVs from high-throughput array-based technologies. There is also a strong interest in identifying associations between CNVs and biological functions. These analyses are commonly done in two stages, by first inferring the CNV calls and using them in the analysis as if they were the true copy numbers. Another commonly used procedure performs the association analysis using the normalized raw intensity measurements. These approaches have several limitations. Here, we propose a hierarchical Bayesian model that handles both the CNV detection and association analysis in a unified manner, by integrating array CGH and gene expression data collected on the same set of subjects. We specify a measurement error model that relates the gene expression levels to the latent copy number states, which in turn are related to the observed surrogate fluorescence intensity measurements via a hidden Markov model. Latent selection indicators that exploit the dependencies between copy number states at adjacent chromosomal locations are incorporated into the model. Model fitting and posterior inference are accomplished via MCMC stochastic search techniques. We demonstrate the performance of the method on simulated data and illustrate its application on a genomic study.

Mahlet G. Tadesse

Georgetown University, Washington, DC, USA. e-mail: mgt26@georgetown.edu

Alberto Cassese

Rice University, Houston, TX, USA. e-mail: alberto.cassese@rice.edu

Michele Guindani

MD Anderson Cancer Center, Houston, TX, USA. e-mail: mguindani@mdanderson.org

Francesco Falciani

University of Liverpool, Liverpool, UK. e-mail: F.Falciani@liverpool.ac.uk

Marina Vannucci

Rice University, Houston, TX, USA. e-mail: marina@rice.edu

Abstract *Le variazioni nel numero di copie (CNV) sono aberrazioni cromosomiche che portano ad un numero di copie anormale nei segmenti di DNA. Il problema dell'identificazione di tali variazioni ha ricevuto grande attenzione, tanto che sono stati proposti molti metodi per la loro identificazione sulla base di dati derivanti da tecnologie basate sugli arrays. Di grande interesse è anche l'identificazione delle associazioni fra le variazioni nel numero di copie e le funzioni biologiche. Tali analisi vengono, di solito, fatte in due fasi: innanzitutto si identificano gli stati latenti delle CNV e si usano, poi, questi ultimi nelle analisi, come se fossero il vero numero di copie. Un'altra procedura, usata comunemente, si basa sull'analisi di associazioni utilizzando misure di intensità grezze normalizzate. Tali approcci hanno molte limitazioni. In questo lavoro proponiamo un modello Bayesiano gerarchico che gestisce simultaneamente l'identificazione delle variazioni nel numero di copie e l'analisi di associazione, tramite l'integrazione dei dati array CGH e di espressione genica raccolti sugli stessi individui. Abbiamo definito un modello con errore di misura, che collega i livelli di espressione genica al numero di copie latenti i quali, a loro volta, sono legati, tramite un hidden Markov model, alle misurazioni di intensità di fluorescenza. Nel modello sono incorporati indicatori latenti di selezione, che sfruttano la dipendenza esistente tra gli stati latenti del numero di copie fra due locazioni cromosomiche adiacenti. L'adattamento del modello e l'inferenza a posteriori sono realizzate attraverso l'utilizzo di tecniche MCMC di ricerca stocastica. Il buon funzionamento del modello è dimostrato su dati simulati e viene, inoltre, presentata un'applicazione su di uno studio di genomica.*

Key words: Bayesian hierarchical models, copy number variation, gene expression, hidden Markov models, measurement error, variable selection

1 Introduction

Copy number variants (CNV) represent DNA fragments having an abnormal number of copies. It is believed that some CNVs have no obvious phenotypic consequence or are merely related to normal phenotypic variations, while others may influence phenotypic outcomes.

The problem of detecting CNVs has received a lot of attention. Several methods, most of which rely on hidden Markov models (HMM) [4, 10], have been proposed. Another question of interest is the identification of CNVs associated with biological functions and complex human diseases. Commonly used procedures use the normalized continuous fluorescence intensity measurements without inferring copy numbers and assess their association with the outcomes of interest using univariate tests or simple linear regression models with multiple testing correction [10]. More recently, multivariate methods have been developed, as in Monni and Tadesse [6] who proposed a stochastic partitioning method to identify sets of correlated gene expression levels and sets of chromosomal aberrations that jointly modulate mRNA transcript abundance in the co-expressed genes. Using the raw measurements has

the advantage of bypassing the need to infer copy number states, but the high noise in the signal intensities may lead to the identification of a large number of false positives [2]. The other widely used approach performs the analysis in two steps by first estimating the copy number states then using these as true states in the association analysis. Determining the underlying CNV state provides insight into the biological mechanisms, but using the estimated copy numbers as if they were the true states ignores the uncertainty in the estimation process and can introduce bias. Methods that incorporate the uncertainty in copy number estimation into the association analysis have recently been proposed [1, 9].

In this work, we present a unified method to simultaneously infer CNV and identify significant associations between copy number states and gene expression levels. This is achieved by specifying a joint distribution of the gene expression and CGH data, which is factored into the product of conditionally independent submodels: an outcome model that relates the gene expression levels to latent copy number states, and a measurement model that relates these latent states to the observed surrogate CGH measurements using a first order HMM. Latent variable selection indicators are introduced to identify CNVs associated with gene expression levels, and selection priors that account for the spatial dependence between adjacent DNA segments are specified. Further details of the model, its implementation, and its performance on simulated and real data can be found in Cassese et al. [3].

2 Model Formulation

Let Y_{ig} denote the expression measurement for gene g ($g = 1, \dots, G$) and X_{im} be the normalized $\log_2$ fluorescence intensity ratio for the m -th CGH probe ($m = 1, \dots, M$) in sample i ($i = 1, \dots, n$). The M CGH probes are considered ordered according to their chromosomal locations, and probes m and $m + 1$ are viewed as being “adjacent”. The observed CGH intensities, X_{im} , are treated as surrogates for the unobserved copy number state, ξ_{im} , which can be in one of four possible states (1=loss, 2=neutral, 3=gain, 4=multiple gains).

2.1 Measurement error model via HMM

We consider a nondifferential measurement error, such that conditional on the latent state ξ , the observed surrogate contains no additional information on the response Y , i.e., $f(Y|\xi, X) = f(Y|\xi)$ [8].

The outcome model relating the expression level of gene g to the latent copy number states ξ is specified using a linear regression model

$$Y_{ig} = \mu_g + \xi_i \beta_g + \varepsilon_{ig}, \quad \varepsilon_{ig} \sim \mathcal{N}(0, \sigma_g^2) \quad i = 1, \dots, n \quad g = 1, \dots, G \quad (1)$$

The measurement model relating the surrogate $\mathbf{X}$ to the latent $\boldsymbol{\xi}$ is defined in terms of the emission distributions of a HMM

$$X_{im} | (\xi_{im} = j) \stackrel{iid}{\sim} \mathcal{N}(\eta_j, \sigma_j^2) \quad j = 1, \dots, 4 \quad (2)$$

and the dependence between the states of adjacent probes is captured by a first order Markov model

$$P(\xi_{i(m+1)} | \xi_{i1}, \dots, \xi_{im}) = P(\xi_{i(m+1)} | \xi_{im}) = a_{\xi_{im} \xi_{i(m+1)}} \quad (3)$$

where $\mathbf{A} = (a_{hj})$ is the matrix of transition probabilities ($h, j = 1, \dots, 4$). The initial probabilities for each state are assumed equal to the stationary distribution π_A .

2.2 Prior specification with spatial dependence

In order to identify the CGH aberrations that affect gene expression levels, we introduce a $(G \times M)$ matrix $\mathbf{R}$ of binary elements, r_{gm} , that are used to specify a spike-and-slab prior on the regression coefficients [5]

$$\beta_{gm} | r_{gm}, \sigma_g^2 \sim r_{gm} \cdot \mathcal{N}(0, c_\beta \sigma_g^2) + (1 - r_{gm}) \cdot \delta_0 \quad (4)$$

where δ_0 is a point mass at 0. We take conjugate normal-gamma priors for the intercepts, μ_g , and the residual precisions, σ_g^{-2} .

Contiguous regions with the same non-neutral copy number state are likely to correspond to the same DNA aberration, and thus likely to jointly affect the expression level of a gene. We, therefore, assume *a priori* that the selection of a probe at location m depends on the copy number states and the selection status of its adjacent probes $\{m-1, m+1\}$. We represent this dependent association structure as follows:

$$\begin{aligned} \pi(r_{gm} | r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}, \pi_1) &= \gamma_m [\pi_1^{r_{gm}} (1 - \pi_1)^{1-r_{gm}}] + \sum_{j=1}^2 \omega_m^{(j)} I_{\{r_{gm}=r_{g(m+(-1)j)}\}} \\ \gamma_m &= \frac{\alpha}{\alpha + s_{(m-1)m} + s_{m(m+1)}} & \pi_1 &\sim \text{Beta}(e, f) \\ \omega_m^{(1)} &= \frac{s_{(m-1)m}}{\alpha + s_{(m-1)m} + s_{m(m+1)}} & \omega_m^{(2)} &= \frac{s_{m(m+1)}}{\alpha + s_{(m-1)m} + s_{m(m+1)}} \end{aligned} \quad (5)$$

$$s_{(m-1)m} = \frac{1}{n} \sum_{i=1}^n \frac{e^{1-\frac{d_m}{D}} - 1}{e - 1} I_{\{\xi_{im} = \xi_{i(m-1)}\}}$$

where d_m is the distance between probes $\{m-1, m\}$, D is the total length of the chromosome, and $\omega_1^{(j)} = \omega_M^{(j)} = 0$ ($j = 1, 2$). The hyperparameter α is a positive scalar that controls the relative strength of the dependence between adjacent probes. As $\alpha \rightarrow \infty$, $\gamma_m \rightarrow 1$, which corresponds to the independent prior.

We complete the prior specification by defining Dirichlet priors for the rows of the transition matrix $\mathbf{A}$, and truncated normal-gamma priors for the mean and precision parameters of the emission distributions.

2.3 Posterior inference

Our main interest lies in the estimation of the association matrix, $\mathbf{R}$, and the copy number states, ξ . After integrating out $\mu_g, \beta_g, \sigma_g^2$, the marginal likelihood reduces to a t -distribution. The model is fit via MCMC by iterating the following steps:

1. Update $\mathbf{R}$ using a Metropolis algorithm by randomly selecting n_g genes and proposing for each gene a change in the inclusion status of a probe using an add/delete/swap move.
2. Update ξ using a Metropolis-Hastings algorithm by randomly choosing a column and proposing new states for a subset of its elements using the current values of the transition matrix.
3. Update the emission distribution parameters, η_j and σ_j^2 , using Gibbs sampling.
4. Update the transition probability matrix, $\mathbf{A}$, using a Metropolis algorithm.

Inference for the elements of $\mathbf{R}$ is performed by calculating their marginal posterior probabilities of inclusion (PPI) from the MCMC output. Those with marginal PPI passing a Bayesian false discovery rate (FDR) threshold are selected as having a significant association [7]. For each individual i , the copy number state for probe m is estimated by the mode of the posterior distribution of ξ_{im} .

3 Simulation Studies

We studied the performance of the model on various simulated datasets. We present the results from two scenarios: one with no spatial dependence structure in the association between markers and genes; the second with dependence created by making clusters of adjacent probes associated with the same gene expression. The model successfully recovered the true gene-CNV associations. Overall, lower α values lead to less false negatives and more false positives. For scenario 1, stronger *a priori* dependence structures (i.e., smaller α) result in an increased number of erroneous decisions, since such structures are not present in the data. For scenario 2, dependent priors provide fewer erroneous decisions and the independent prior ($\alpha = \infty$) shows the worst performance, since it does not use the information from adjacent probes. As a general guideline, moderate values of α give an appropriate compromise between false positives and false negatives.

We compared the results from our proposed unified method to single stage approaches that focus solely on CNV detection or solely on association analysis using the raw measurements. Our method yielded fewer misclassified CNV calls com-

pared to the former and it detected substantially fewer spurious associations compared to the latter.

4 Application to NCI-60 data

We applied the method to the NCI-60 cancer cell line panel using pathway-based scores for the gene expression data and focusing on CGH probes mapping to chromosome 8. Our results identified potential links between CNVs and the transcriptional activity of target pathways. Four pathway components were selected as those with the highest number of associations. In particular, connections of six genetic mutations with Arginine metabolism, and 18 genetic mutations with the expression of genes involved in Glycosylphosphatidylinositol anchor metabolism and Porphyrin metabolism were identified. We searched for putative upstream regulators of all genes involved in these pathways and of genes corresponding to the selected CNVs, and found four putative upstream regulators that consisted of the well-known oncogenes MYC and p53, PPAR and SOD1. These findings support the hypothesis that the associations we have identified may represent genes implicated in cancer.

References

1. Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J. et al.: A robust statistical method for case-control association testing with copy number variation. *Nat. Gen.* **40** 1245-1252 (2008)
2. Breheny, P., Chalise, P., Batzler, A., Wang, L., Fridley, B.L.: Genetic association studies of copy number variation: Should assignment of copy number states precede testing? *PLoS ONE*. **7** e34262 (2012)
3. Cassese, A., Guindani, M., Tadesse, M.G., Falciani, F., Vannucci, M.: A hierarchical Bayesian model for inference of copy number variants and their association to gene expression. *Annals of App. Stat.* in press (2014)
4. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H. et al.: QuantiSNP: an objective Bayes hidden Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35** 2013-2025 (2007)
5. George, E., McCulloch, R.E.: Approaches for Bayesian variable selection. *Stat. Sinica*. **7**, 339-373 (1997)
6. Monni, S., Tadesse, M.G.: A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Anal.* **4**, 413-436 (2009)
7. Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. **5**, 155-176 (2004)
8. Richardson, S., Gilks, W.R.: Conditional independence models for epidemiological studies with covariate measurement error. *Stat. Med.* **12**, 1703-1722 (1993)
9. Subirana, I., Diaz-Uriarte, R., Lucas, G., Gonzalez, J.R.: CNVassoc: Association analysis of CNV data using R. *BMC Med Genomics*. **4** 4-47 (2011)
10. Wang, K., Li, M., Hardley, D., Liu, R., Glessner, J. et al: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection inn whole-genome SNP genotyping data. *Genome Res.* **17** 1665-1674 (2007)