SPATIAL ANALYSIS OF COMPLEX BIOLOGICAL TISSUES FROM SINGLE CELL GENE EXPRESSION DATA

Clustering and visualizing functionnal tissues in *Platynereis dumerillii*

JEAN-BAPTISTE OLIVIER GEORGES PETTIT

UNIVERSITY OF CAMBRIDGE

2014 – version 0.9

[January 1, 2014 at 17:30 – classicthesis version 0.9]

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.
1939 – 2005

ABSTRACT

This is wehere the abstarct will go...

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Put your publications from the thesis here. The packages `multibib` or `bibtopic` etc. can be used to handle multiple different bibliographies in your document.

Le temps ne fais rien a l'affaire...

— George Brassens

ACKNOWLEDGMENTS

My thanks will go there

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

SPATIAL ANALYSIS OF COMPLEX BIOLOGICAL TISSUES FROM SINGLE CELL GENE EXPRESSION DATA

1 INTRODUCTION

This is where the introduction goes

2

ANALYSING COMPLEX BIOLOGICAL TISSUES, THEIR EXPRESSION IN A SPATIAL VISUALIZATION

3

HIDDEN MARKOV RANDOM FIELD BASED CLUSTERING FOR SINGLE CELL GENE EXPRESSION DATA

3.1 MARKOV RANDOM FIELD PRIOR DISTRIBUTION

Let S be a finite set of sites, each of which represents one “cube" of data (see the results section for a detailed description of the data). Given the coordinates of each site, we were able to define a neighbourhood system on S using a first order neighbourhood system, i.e the 6 closest sites. S and its neighbourhood system can be viewed as a connecting graph G. Let C be the set of cliques of G. C is therefore the set of all sites that are all neighbours from one another.

Let a Random Field Z be defined as a set of random variables $Z = \{Z_i, \forall i \in S\}$ each Zi taking its value in $[1, K]$. For every site $i \in S$, let $N(i)$ represent the set of its neighbours and $\mathbf{z}_{S-\{i\}}$ a realization of the field restricted to $S - \{i\} = \{j \in S, j \neq i\}$. Z is a Markov Random Field if and only if it verifies the Markov property at every site :

$$\forall i \in S, P_G(z_i \mid z_{S-\{i\}}) = P_G(z_i \mid z_j, j \in N(i)) \tag{1}$$

Equation (1) states that the realization of the field at any site $i \in S, z_i$ can be fully determined using only the state of its neighbours $N(i)$. In other words each “cube" is only dependent upon its neighbours. The Hammersley-Clifford theorem state that if Z is a Markov Random Field, the join distribution of the field follows a Gibbs distribution so that :

$$P_G(\mathbf{z}) = \frac{e^{-H(\mathbf{z})}}{\sum_{\mathbf{z}'} e^{-H(\mathbf{z}')}} \tag{2}$$

$H(\mathbf{z})$ is called the Energy function and is summed over the cliques of the graph C. Considering that we are working with an order one neighbouring graph, C is the set of all the couples of sites (i, j) that are neighbours. We chose to consider H as a function of vector $\text{ } = (\beta_1, \ldots, \beta_K)$ containing K parameters, one per cluster and $v_{i,j}$ a potential function set to 1 in our method.

$$H(\mathbf{z}) = -\sum_{i \in S} \beta_{z_i} \sum_{\substack{i,j \\ neighbours}} v_{i,j} \times_{[z_i = z_j]} \tag{3}$$

The denominator in (2) where $\mathbf{z}'$ represents all the possible realizations of the field is a normalizing constant that we will refer to as

$W(\zeta)$.
This model is closely related to a K-color Potts model [7] although instead of a single parameter β for the entire model, we assign one β per cluster. Equation (3) is a decreasing function of every component of ζ and of the number of neighbouring "cubes" in the field having the same class. This Energy thus favours spatially regular partitions and a higher value of β_h, with $1 \leqslant h \leqslant K$ will amplify the smoothing effect, or coherence over cluster h. We chose to use one spatial smoothness parameter per cluster because of the nature of the data we are dealing with. Indeed, in a biological context, it is expected that some tissues will be more spatially coherent than others.

From this prior distribution we have K unknown parameters $\zeta = (\beta_1, \ldots, \beta_K)$ to be estimated by the model. It is important to note at this point that $W(\zeta)$ is summed over all possible realizations of the field Z, this is an exponentially complex sum as the cardinality of S rises. Therefore the computation of the normalizing factor becomes intractable very quickly. To address this problem, we are going to need to make some approximations in order to compute this quantity (see Mean Field Approximations).

We have described the prior distribution of a Markov Random Field representing our partition, we now need to describe the relationship between Z and the data.

3.2 HIDDEN MARKOV MODEL

As Z is unknown a priori and represents the partition, let Y be a set of random variables representing the observations (the in-situ hybridization data). We have to assume conditional independence of the observations given the partition Z so that, with f_{z_i} the density function relative to cluster $z_i, i \in S$:

$$p(\mathbf{y} \mid \mathbf{z}; \Theta) = \prod_{i \in S} p(y_i \mid z_i; \Theta) \tag{4}$$

$$= \prod_{i \in S} f_{z_i}(y_i \mid z_i; \Theta) \tag{5}$$

We define one unknown parameter per cluster: $\Theta = (\eta_1, \ldots, \eta_K)$. It is interesting to note that this part of the model is equivalent to an independent mixture model [4]. Indeed, hidden Markov models can be viewed as independent mixture models where Z is a set of independent, identically distributed random variables, which happens when $\beta = 0$.

Because the 169 genes chosen by Tomer et al. [6] are key genes involved in the early development of Platynereis' brain, the assumption

of conditional independence given the realization of the field seems acceptable. The validity of this hypothesis may however be argued for future RNA-seq datasets representing entire transcriptomes (see discussion).
Given a particular cluster $h \in [1, K]$ and M the set of considered genes, we assume that each gene $m \in M$ follows a Bernoulli distribution with parameter $\theta_{h,m}$. We then have one unknown Bernoulli parameter per gene per cluster so that :

$$\begin{aligned}\Theta &= (\alpha_1, \ldots, \alpha_K) \\ &= \begin{pmatrix} \theta_{1,1} & \ldots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \ldots & \theta_{M,K} \end{pmatrix}\end{aligned}$$

The conditional density function $f_i, i \in S$ can be expressed as :

$$f_i(y_i \mid z_i; \Theta) = f_i(y_i \mid z_i; \alpha_{z_i}) = \prod_{m \in M} \theta_{z_i,m}^{y_{i,m}} \times (1 - \theta_{z_i,m}^{1-y_{i,m}}) \qquad (6)$$

Looking at both fields Z and $Y \mid Z$ together, the complete likelihood of the model is expressed as :

$$P_G(\mathbf{y}, \mathbf{z} \mid \Theta, \beta) = f(\mathbf{y} \mid \mathbf{z}, \Theta) P_G(\mathbf{z} \mid \beta) = \frac{exp\{-H(\mathbf{z} \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i, \theta_z i)\}}{\sum_{z'} e^{-H(\mathbf{z})}} \qquad (7)$$

Because equation (7) is a Gibbs distribution, using the Hammersley-Clifford theorem we can conclude that the conditional field Y given $Z = \mathbf{z}$ is another a Markov Random Field with the Energy function

$$H(\mathbf{z} \mid \mathbf{y}, \beta, \Theta) = H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log f_i(y_i \mid z_i, \Theta)$$

In our case, the goal is to recover the unknown realization of $Z : \mathbf{z}$. To this end we need to maximize the values of all the parameters of the model $\Phi = (\Theta, \beta)$. We will also need to determine the unknown value K. This will be determined a posteriori by computing the BIC over the model full likelihood [5].

3.3 PARAMETER ESTIMATION USING THE EM ALGORITHM

The EM principle can be applied to estimate the parameters $\Phi = (\Theta, \beta)$ of the hidden MRF model. After initializing the clusters $\mathbf{z}$, we

choose Φ^{l+1} at iteration $(l+1)$ in order to maximize the model's expectation:

$$\begin{aligned} Q(\Phi \mid \Phi^l) &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{y}, \mathbf{z}; \Phi) \\ &= \underbrace{\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{y} \mid \mathbf{z}; \Theta)}_{R_y(\Theta \mid \Phi^l)} + \underbrace{\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \log p(\mathbf{z} \mid \zeta)}_{R_z(\zeta \mid \Phi^l)} \end{aligned} \tag{8}$$

The decomposition in (8) allows us to consider separately the maximization of $R_y(\Theta \mid \Phi^l)$ and $R_z(\zeta \mid \Phi^l)$:

$$\Theta^{l+1} = \arg\max_{\Theta} R_y(\Theta \mid \Phi^l)$$
$$\zeta^{l+1} = \arg\max_{\zeta} R_z(\zeta \mid \Phi^l)$$

We estimate $Q(\Theta \mid \Phi^l)$ in the E step by further developing it using equation (5):

$$\begin{aligned} Q(\Theta \mid \Phi^l) = R_y(\Theta \mid \Phi^l) &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}; \Phi^l) \sum_{i \in S} \log f_{z_i}(y_i; \Theta) \\ &= \sum_{i \in S} \sum_{h=1}^{K} [\log f_h(y_i; \Theta)] \; p(Z_i = h \mid \mathbf{y}; \Phi^l) \end{aligned}$$

Therefore, at each iteration we need to compute in the E step the following quantity :

$$t_{ih}^{m+1} = p(Z_i = h \mid \mathbf{y}; \Phi^l)$$

Computing this conditional probability is problematic because of the dependence between neighbouring "cubes", and an exact value cannot be obtained without considerable computing resources. As mentioned previously, we also need to approximate the normalizing constant $W(\zeta)$. Approximation methods include Besag's pseudo-likelihood [1] to compute $W(\zeta)$, and simulating the posterior distribution of Z given $\mathbf{y}$ with the parameters at iteration (l), with a Gibbs sampler to estimate t_{ih}^{m+1} [2].

However, another method exists, the mean field approximation originally proposed in the field of statistical mechanics. Since then, it has been used in a variety of fields including computer vision [8] and more recently to approximate the distribution of both $W(\zeta)$ (with a single β) and t_{ih}^{m+1} [9]. We present here the extension of this method to a model with $\zeta = (\beta_1, \ldots, \beta_K)$.

3.4 MEAN FIELD APPROXIMATIONS

The idea behind this approximation is to compute intractable quantities at any point $i \in S$ by setting the values of all the other sites in

the field to their mean values. As seen in equation (1) in the case of a MRF, this is equivalent to fixing the values of $N(i)$ only.

When computing t_{ih}^{m+1}, the mean fields approximation yields the following fixed point equation [3] for $i \in S$ and $1 \leqslant h \leqslant K$:

$$t_{ih}^{m+1} \approx \frac{f_h(y_i; \nu_h^m) \exp\{\beta_h^m \sum_{j \in N(i)} t_{jh}^{m+1}\}}{\sum_{u=1}^{K} f_u(y_i; \nu_u^m) \exp\{\beta_u^m \sum_{j \in N(i)} t_{ju}^{m+1}\}} \tag{9}$$

For the normalizing constant $W(\zeta)$, if we apply the mean-field approximation, using equation (3), we can write :

$$W(\zeta) = \sum_{\mathbf{z}'} \exp(-H(\mathbf{z}')) \approx \sum_{i \in S} \sum_{\mathbf{z_i}} \exp(-H(\mathbf{z_i})) = \sum_{i \in S} \sum_{\mathbf{z_i}} \exp(\beta_{z_i} \sum_{N(i)} [z_i = z_j])$$

With this new set of equations, we are now able to estimate all quantities needed in the E step to maximize the model's expectation.

3.5 MAXIMIZATION

After the E step, the maximizing Φ is relatively straight forward. For Θ, once the the $t_{ih}^m = p(Z_i = h \mid \mathbf{y}; \Phi^l)$ have been computed during the E-step, we use those probabilities to assign each cell to its cluster at step l. Once the new partition is created, the maximization of Θ can be computed iteratively for cluster $h \in [1, K]$ and gene $m \in M$ with $Expr_{h,m}$ the number of cells expressing gene m in cluster h and Num_h the total number of cells in cluster h.

$$\theta_{m,h}^{l+1} = \arg\max_{\Theta} R_y(\Theta \mid \Phi^l) = \frac{Expr_{h,m}}{Num_h}$$

We then need to maximize ζ^{l+1}, to this end, we use a gradient ascent algorithm for each $\beta_h^{l+1}, h \in [1, K]$ on the function $R_z(\zeta \mid \Phi^l)$. This process is done iteratively. (CITE LAMIAE)

We are now able to compute a partition over K clusters by applying the previously described EM algorithm. However, we still need to find a way of choosing K.

3.6 ESTIMATING K

Without any prior knowledge, choosing the right number of clusters K is challenging. We decided to use an a posteriori method relying on the final log Likelihood of the model derived from equation (7):

$$\log L(\Phi) = \log P_G(\mathbf{y}, \mathbf{z} \mid \Theta, \zeta)$$

Because $\mathrm{Log}\ L(\Phi)$ monotonically increases with the number of parameters of the model, the BIC approach penalizes the addition of new parameters to the model. Let P be the total number of parameters in the model and N the cardinality of S, the BIC is expressed as:

$$-2 \log L(\Phi) + P \log N$$

By computing the final likelihood for a large range of possible K values, the minimal resulting BIC will be chosen as the optimal number of classes, $\hat{K}$ for our dataset. This approach is not ideal (see Discussion) but yields good results when applied to simulated data (see Results).

Part II

APPENDIX

A

APPENDIX

BIBLIOGRAPHY

[1] Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.

[2] Bernard Chalmond. An iterative gibbsian technique for reconstruction of< i> m</i>-ary images. *Pattern recognition*, 22(6):747–761, 1989.

[3] M. Dang and G. Govaert. Spatial fuzzy clustering using EM and markov random fields. *International Journal of System Research and Information Science*, 8(4):183–202, 1998.

[4] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley. com, 2004.

[5] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[6] Raju Tomer, Alexandru S Denes, Kristin Tessmar-Raible, and Detlev Arendt. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5):800–809, 2010.

[7] Fa-Yueh Wu. The potts model. *Reviews of modern physics*, 54(1): 235, 1982.

[8] Alan L Yuille. Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, 2(1):1–24, 1990.

[9] Jun Zhang. The mean field theory in em procedures for markov random fields. *Signal Processing, IEEE Transactions on*, 40(10):2570–2583, 1992.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

DECLARATION

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specied in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualication at any other university; and
- does not exceed the prescribed limit of 60,000 words.

Cambridge, 2014

Jean-Baptiste Olivier
Georges Pettit