Clustering of longitudinal shape data sets using mixture of separate or branching trajectories

Vianney Debavelaere, Alexandre Bône, Stanley Durrleman, Stéphanie Allassonnière

Clustering of longitudinal shape data sets using mixture of separate or branching trajectories

Vianney Debavelaere[1], Alexandre Bône[2], Stanley Durrleman[2], and Stéphanie Allassonnière[3], for the Alzheimers Disease Neuroimaging Initiative

[1] CMAP, Ecole Polytechnique
[2] CRC, Université Paris Descartes
[3] INRIA, Aramis team, ICM institute

Abstract. Several methods have been proposed recently to learn spatiotemporal models of shape progression from repeated observations of several subjects over time, i.e. a longitudinal data set. These methods summarize the population by a single common trajectory in a supervised manner. In this paper, we propose to extend such approaches to an unsupervised setting where a longitudinal data set is automatically clustered in different classes without labels. Our method learns for each cluster an average shape trajectory (or representative curve) and its variance in space and time. Representative trajectories are built as the combination of pieces of curves. This mixture model is flexible enough to handle independent trajectories for each cluster as well as fork and merge scenarios. The estimation of such non linear mixture models in high dimension is known to be difficult because of the trapping states effect that hampers the optimisation of cluster assignments during training. We address this issue by using a tempered version of the stochastic EM algorithm. Finally, we apply our algorithm on synthetic data to validate that a tempered scheme achieve better convergence. We show then how the method can be used to test different scenarios of hippocampus atrophy in ageing by using an heteregenous population of normal ageing individuals and mild cognitive impaired subjects.

Keywords: Longitudinal data analysis · Mixture model · Branching population · Riemannian framework.

1 Introduction

With the emergence of large longitudinal data sets (subjects observed repeatedly at different time points), construction of spatiotemporal atlases has become a central issue. From the repeated observations of individuals at different time points, such atlases aim at estimating a trajectory that will be representative of the population, as well as the spatiotemporal variability within this population. The representative trajectory is a long-term scenario of changes informed by sequences of short-term individual data. Being able to construct such an atlas has a large number of applications: understanding of a disease progression, highlighting of growth patterns, etc.. Several statistical approaches have been proposed

to address this problem. Descriptive [7] or generative [10,13] approaches have been derived for data taking the form of feature vectors. Generative longitudinal models have been proposed for shape data, usually using flows of deformations to construct shape trajectories [5,11,14]. All these methods however assumed that observations are drawn from an homogeneous population that may be summarized by a single representative trajectory.

In many situations, populations are likely to be heterogeneous but without prior knowledge on the sub-groups composing them, thus preventing the use of such supervised approaches. Developing unsupervised statistical learning methods is known to be challenging. The difficulty is further increased in the spatiotemporal setting since clustering may take various forms: sub-groups may follow independent trajectories, or they may follow trajectories that fork or merge at specific time-points. The former case is relevant to discover pathological sub-types having different disease course. The latter is interesting for a disease that is seen as a progressive deviation from a normal aging scenario.

In this paper, we address this issue by proposing a mixture model for longitudinal shape data. The scenario of each cluster results from the combination of successive pieces of trajectories. This framework enables to build complex broken line design. These pieces of trajectory may be different for different clusters, or certain parts may be shared among clusters to model forking or merging trajectories. The model offers therefore a very flexible framework for testing complex clustering scenarios.

In practice, we used the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework to build trajectories of shape changes, which may be seen as geodesics on a Riemannian manifold [12]. The construction of longitudinal atlases has been proposed in this framework, where the representative trajectory is a geodesic [5,13] or a piecewise geodesic [2] and the spatiotemporal variability is modeled using a tubular coordinate system around the common geodesic. Such methods are based on a generative mixed-effect model. We extend here this model to a mixture model, where each mixture component is described by a piecewise geodesic curve. Some part of this curve may be shared by several clusters. To estimate the different parameters, derivations of the Expectation-Maximization algorithm are known to be efficient [4]. However, estimating mixture components in such a high-dimension non-linear setting is known to be difficult. A central difficulty is the "trapping states" effect, where changing class assignment is always more costly than adjusting the parameters of the clusters, resulting in very few updates of class assignment during optimization. Pragmatic solutions have been proposed, for instance in [3] for cross sectional data but at high computational cost. Here we introduce tempered distributions into the stochastic approximation EM in order to avoid to be trapped in the initial labelling.

We will finally quantitatively validate our model on simulated 2D data and then apply it on a data set of hippocampus shapes in the context of Alzheimer's disease to highlight different atrophy patterns between normal ageing individuals and subjects developing Alzheimer's disease.

2   Geometrical model

In the following, we consider a longitudinal data set of n subjects, each being observed k_i times: $(y_{i,j})_{1\leq i\leq n,1\leq j\leq k_i}$ at time $(t_{i,j})_{1\leq i\leq n,1\leq j\leq k_i}$.

We recall the construction of geodesics in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [9]. Given an initial set of control points c_0 and momenta m_0 at a time t_0, we obtain a flow of diffeomorphisms $Exp_{c_0,t_0,t}(m_0)$ that deforms the ambient space continuously in time, and then any mesh embedded into this space. These flows define geodesics on the manifold $M = \left\{ Exp_{c_0,t_0,1}(m_0) | m_0 \in \mathbb{R}^{n_{cp}\times d} \right\}$.

We want now to construct a piecewise geodesic representative trajectory γ_0 as a combination of K different geodesics following each other, generalizing the work done in [2] in dimension 1. Hence, we introduce a subdivision of $\mathbb{R}$: $(t_{R,1} < ... < t_{R,K-1} < t_{R,K} := +\infty)$ where $(t_{R,k})_{1\leq k\leq K-1}$ are the rupture times i.e. times when the representative curve switches from one geodesic to the other. We also fix the value of the representative curve at the rupture times y_k for $k \geq 2$ to assure the continuity of the trajectory. More formally, given a set of control points $c_1 \in \mathbb{R}^{n_{cp}\times d}$, of rupture times $t_R \in \mathbb{R}^{K-1}$, an initial shape y_1 and K momenta $(m_0, m_1, ..., m_{K-1})$, we define the representative trajectory as:

$$
\begin{cases}
\gamma(t)(y_1) = \mathrm{Exp}_{c_1,t_{R,1},t_{R,1}-t}(m_0) \circ y_1 \mathbb{1}_{t\leq t_{R,1}} + \sum_{k=1}^{K-1} \mathrm{Exp}_{c_k,t_{R,k},t-t_{R,k}}(m_k) \circ y_k \mathbb{1}_{t_{R,k}\leq t\leq t_{R,k+1}} \\
\qquad c_k = \mathrm{Exp}_{c_{k-1},t_{R,k-1},t_{R,k}-t_{R,k-1}}(m_{k-1}) \circ c_{k-1} \qquad \text{for } k \geq 2 \\
\qquad y_k = \mathrm{Exp}_{c_{k-1},t_{R,k-1},t_{R,k}-t_{R,k-1}}(m_{k-1}) \circ y_{k-1} \qquad \text{for } k \geq 2.
\end{cases}
$$

$$(1)$$

It can be remarked that the first rupture time has a particular role as we must define a geodesic before it and another after.

Now that we can define a representative trajectory, we want to compute a transformation from this trajectory towards a subject taking into account both spatial and temporal differences. For each subject i, let $\xi_{i,0}, ...\xi_{i,K-1}$ be acceleration coefficients and $\tau_{i,0}, ..., \tau_{i,K-1}$ time shifts. We write for every subject i: $\psi_{i,0}(t) = t_{R,1} - e^{\xi_{i,0}}(t_{R,1} - t + \tau_{i,0})$ and, for each component $k \geq 1$, $\psi_{i,k}(t) = t_{R,k} + e^{\xi_{i,k}}(t - t_{R,k} - \tau_{i,k})$. Each subject has its own rupture times: $t_{R,i,k}$ such that $t_{R,k} = \psi_{i,k}(t_{R,i,k})$ i.e. $t_{R,i,k} = t_{R,k} + \tau_{i,k}$. To assure the continuity of the time reparametrization at each of those rupture times, we also fix all the time shifts but $\tau_{i,0}$, from now on noted τ_i, by continuity conditions. Finally, we set: $\psi_i(t) = \psi_{i,0}(t)\mathbb{1}_{t\leq t_{R,i,1}} + \sum_{k=1}^{K-1}\psi_{i,k}(t)\mathbb{1}_{t_{R,i,k}\leq t\leq t_{R,i,k+1}}$. The time shifts τ_i allow the subjects to be at different stage of evolution while the acceleration factors $\xi_{i,k}$ allow an inter-subject variability in the pace of evolution on each geodesic (quicker evolution if $\xi_{i,k} > 0$, slower if $\xi_{i,k} < 0$).

As proposed in [5], we also introduce for each subject a space-shift momenta w_i which accounts for geometric variability. We use the notion of parallel transport to define the spatial deformation at a time t. More precisely, we note P_γ the parallel transport which transports any vector $w \in \mathbb{R}^{n_{cp}\times d}$ along the trajectory γ. Then, to code the deformation field at a time t, we transport the
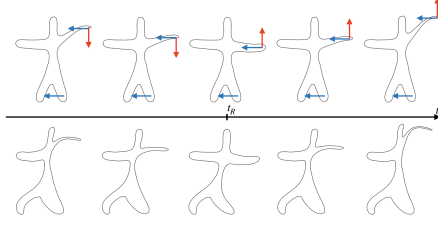
Fig. 1: Samples from a piecewise geodesic (top) and a parallel deformation (bottom). The red momenta codes the template evolution, lowering then raising an arm after the rupture time. The blue momenta is transported along the piecewise geodesic and defines the deformation frame.

momenta w along the curve $\gamma(t)$ and then compute the flow given by this new momenta. More precisely, we define: $\eta_t(w) = \mathrm{Exp}_{\gamma(t)(c_1),0,1}(P_{\gamma(t)}(w))$. Finally, the deformation of the representative curve γ by the space shift w is given by: $\gamma_w(t) = \eta_t(w) \circ \gamma(t) \circ y_1$. The space deformation process is summarized in Fig 1. We will model this space shift as a linear combination of n_s sources: we suppose that $w = A_{(m_0,...,m_{K-1})^\perp}s$ with $A_{(m_0,...,m_1)^\perp}$ a $n_{cp} \times n_s$ matrix called the modulation matrix and $s \in \mathbb{R}^{n_s}$ the sources. By projecting all the columns of $A_{(m_0,...,m_{K-1})^\perp}$ on $(m_0,...,m_{K-1})^\perp$ for the metric K_g, we impose orthogonality between the space shifts and the momenta vectors. It has been shown in [13] that this condition is necessary to assure the identifiability of the model by preventing a confusion between the space shifts and the acceleration factors. Finally, we deform the template $\gamma(t)(y_1)$ by setting: $\gamma_i(t) = \gamma_w(\psi_i(t))$.

This construction builds a piecewise geodesic model of progression. We now propose an extension for the analysis of heterogeneous populations. More precisely, we suppose it exists N different representative curves, each of the subjects i being in the cluster $cl(i)$ of the particular representative curve $\gamma^{cl(i)}$. This representative curve comes with its own set of rupture times $(t_{R,1}^{cl(i)} < ... < t_{R,K-1}^{cl(i)})$, initial shape $y_1^{cl(i)}$, control points $c_1^{cl(i)}$, momenta $(m_0^{cl(i)},...,m_{K-1}^{cl(i)})$ and modulation matrix $A_{(m_0^{cl(i)},...,m_{K-1}^{cl(i)})^\perp}^{cl(i)}$.

This mixture framework enables to compare and test hypothesis on the clusters. For instance, all or part of the clusters can be shared on the first stage of evolution ($t < t_{R,1}$). This imposes all or part of the representative curves to have the same first component, i.e. $t_{R,1}^k$, y_1^k, c_1^k and m_0^k are equals for all or part of the clusters $k \in [|1,N|]$. Similarly, we could impose the equality of some of the representative curves on other time segments, allowing us to handle populations forking or merging at the rupture times.

3   Statistical Model and estimation

We now define the mixed effects statistical model: we note $z_{pop}^r = ((m_k^r)_{0 \leq k \leq K-1}, y_1^r, c_1^r, (t_{R,k}^r)_{0 \leq k \leq K-1}, A_{(m_0^r,...,m_{K-1}^r)^\perp}^r)$ the population parameters of the cluster r and $z_i = ((\xi_{i,k})_{1 \leq k \leq K-1}, \tau_i, s_i)$ the deformation parameters of the subject i. Depending on the case, we place ourselves in the current [15] or var-

ifold [6] framework, allowing us to compute distances between shapes without any point correspondence. We suppose that the subject i is obtained as a noised deformation of the representative curve $\gamma^{cl(i)}$: $\forall i \in [|1,n|]$, $\forall j \in [|1,k_i|]$, $y_{i,j}|cl(i), z_{pop}^{cl(i)}, z_i \sim \mathcal{N}(\gamma_i(t_{i,j}), \sigma^2 Id)$, that the deformation parameters z_i verify: $z_i|cl(i) \sim \mathcal{N}(0, \Sigma^{cl(i)})$ where for all cluster r, Σ^r is a positive-definite matrix, that the cluster r is drawn with a probability p^r i.e. $cl(i) \sim \sum_{r=1}^{N} p^r \delta_r$ and that $z_{pop}^r \sim \mathcal{N}(\bar{z}_{pop}^r, v_{pop}^r)$ where v_{pop}^r are small fixed variances. Our model is thus defined with parameters $\theta = \left((\Sigma^r, p^r, \bar{z}_{pop}^r)_{1 \leq r \leq N}, \sigma \right)$. For effectiveness in the high dimension low sample size setting, we work in the Bayesian framework and set as priors: $\Sigma_r \sim \mathcal{W}^{-1}(V, m_\Sigma)$, $\sigma \sim \mathcal{W}^{-1}(v, m_\sigma)$, $p \sim \mathcal{D}(\alpha)$ and $\bar{z}_{pop}^r \sim \mathcal{N}(\bar{\bar{z}}_{pop}^r, \bar{v}_{pop}^r)$ where $\mathcal{W}$ is the inverse Wishart distribution and $\mathcal{D}$ is the Dirichlet distribution. We can remark that the joint distribution is in the curved exponential family which guaranties the convergence of the Stochastic Approximation Expectation Maximization algorithm (SAEM) [4].

From now on, we write q the probability distribution. To estimate the parameters θ, we want to compute a maximum a posteriori estimator by using a stochastic version of the Expectation Maximization algorithm known as MCMC-SAEM [3]. It consists in the following steps: (i) simulation of (z, z_{pop}, cl) as an iterate of an ergodic Monte Carlo Markov Chain with stationary distribution $q(z_{pop}, z, cl|y, \theta)$, (ii) stochastic approximation of the sufficient statistics of the curved exponential model and (iii) maximization using the updated stochastic approximation.

However, using the algorithm as presented above yields to bad results in exploring the support of the conditional probability distribution. This issue is known as trapping states: once a label is given to an observation, the probability of changing to another is almost zero. This leads to no change of label after a few iterations. To solve this problem, we use a tempered version of the MCMC-SAEM. We choose to sample from $\frac{1}{C(T_k)} q(c|y, \theta)^{\frac{1}{T_k}}$ where T_k is a sequence of temperature converging towards 1 and $C(T_k)$ is the normalizing constant. The higher the temperature, the flatter the distribution and the more the clusters are likely to explore the entire set. We optimize the temperature sequence so that the Markov chain strides well the support with a dynamic similar to a simulated annealing [8]. By choosing a sequence of temperature converging towards 1, we do not affect the convergence of our algorithm, as showed in [1].

4   Results

We first test our algorithm on simulated data. We create 200 subjects by deforming a branching piecewise-geodesic representative curve with two components. More precisely, we draw three random momenta and apply them on a fixed shape to obtain the first common component and the two distinct ones forking at the rupture time. We then apply our algorithm to find the two clusters, the representative curves and the spatiotemporal deformations towards the data sequence of each subject. Results in Fig. 2 show that there is no noticeable differences between the true and estimated trajectories (left), nor between true and
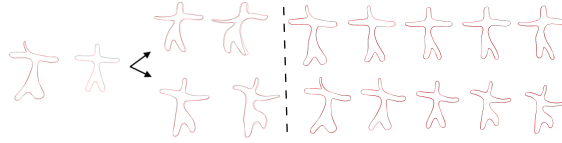
Fig. 2: In red, the exact simulated data, in black, the results given by our algorithm. On the left, the representative curves that split up at a certain rupture time. On the right side, two subjects given with their reconstructions.
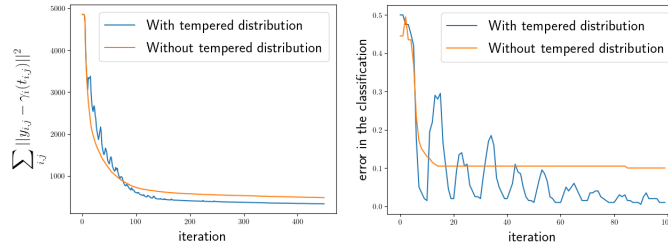

Fig. 3: Left: evolution of the varifold distances between the subjects and their reconstructions. Right: percentage of error in the classification along the first 100 iterations. With tempered distribution, the oscillating temperature coerces a lot of subjects to change classes. After 500 iterations, the error is 31.3% smaller.

reconstructed observations (right). To quantify the reconstruction error, we compute the varifold norm of the errors for all subjects along the iterations on Fig. 3. In particular, all the errors in the estimation of the population and subjects parameters are under 5%. We also show the necessity of using tempered distributions by plotting the error of classification with and without temperature.

We want now to test hypothesis about the heterogeneity of the population. We run our algorithm supposing first that the two representative trajectories are different. We then run it again supposing that their first component is the same and that they fork at the rupture time. To select the model, we then compute the log-likelihood ratio test. As expected, the log-likelihood is smaller for the model without fork (difference of 24.6), confirming that the data are more likely to be drawn from a branching model than two unrelated components.

We now test our algorithm on 100 subjects obtained from the Alzheimer's Disease Neuroimaging Initiative database. 50 of those subjects are control patients (CN) and 50 are Mild Cognitive Impairment subjects eventually diagnosed with Alzheimer's disease (MCIc). Meshes of the right hippocampus is segmented from the rigidly registered MRI. We run our algorithm with a forking model as in the synthetic experiment. As there is no reason for the control subjects to have two different dynamics, we also ask one of the branch to follow the same geodesic after the rupture time. Our algorithm splits the patients in two clusters, one of them presenting a quicker and different pattern of atrophy (Fig. 4 and 5). In particular, 72% of the subjects are classified as expected: the CN in the
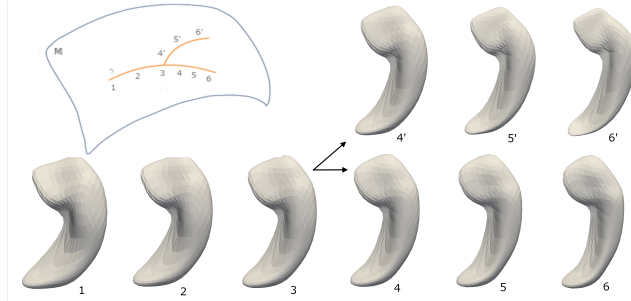
Fig. 4: Representative shape evolution at the ages 63.6y, 68.4y, 74.7y (i.e. rupture time), 77.9y, 81.5y and 85.1y. Bottom shapes: cluster with one dynamic. Top shapes : cluster with change of dynamic after rupture time.
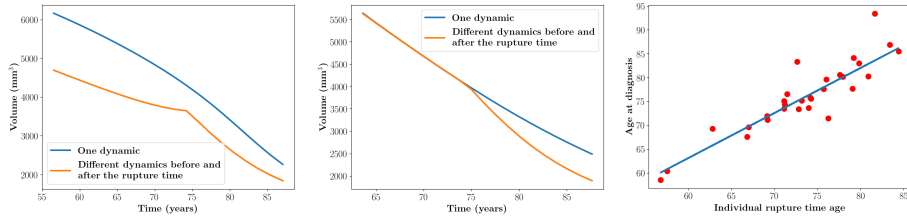

Fig. 5: Left: volume evolution for two clusters with separate trajectories. Center: volume evolution for two branching clusters. Right: comparison of the age at diagnosis with the individual rupture time for the MCIc patients, $R^2 = 0.91$

cluster with one dynamic and a slower atrophy and the MCIc in the cluster with a quicker atrophy after the rupture time. Moreover, the individual rupture times are strongly correlated to the diagnostic age.

We run again the algorithm, looking for two clusters with separate trajectories, one of them with only one dynamic (see supplementary material and Fig 5). This time, 70% of the subjects are classified as expected. It must be remarked that we cannot compare the volume evolution of the different models as the time reparametrization and so the time-line is different from one cluster to the other. We now want to select the model. The log-likelihood is bigger in this second case, which suggests that the MCI subjects are likely to deviate from a normal aging scenario at an earlier stage in life that is not observed in this data set.

5   Conclusion

We proposed a mixture model for longitudinal shape data sets where representative trajectories take the form of piecewise geodesic curves. Our model can be applied in a wide variety of situations to test whether sub-populations fork or merge at different time-points or follow unrelated trajectories on other time

intervals. We showed on simulated examples that our tempered optimisation scheme is key to achieve convergence of such a mixed model combining discrete variables with continuous variables of high dimension. Its application on real data allowed us to investigate the relationship between normal and pathological ageing. Future work will focus on designing specific model selection criterion in this longitudinal setting.

References

1. Allassonnière, S., Chevallier, J.: A new class of em algorithms. escaping local minima and handling intractable sampling (2019)
2. Allassonniere, S., Chevallier, J., Oudard, S.: Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In: Advances in Neural Information Processing Systems. pp. 1152–1160 (2017)
3. Allassonnière, S., Kuhn, E.: Stochastic algorithm for bayesian mixture effect template estimation. ESAIM: Probability and Statistics **14**, 382–408 (2010)
4. Allassonnière, S., Kuhn, E., Trouvé, A., et al.: Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. Bernoulli **16**(3), 641–678 (2010)
5. Bône, A., Colliot, O., Durrleman, S.: Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9271–9280 (2018)
6. Charon, N., Trouvé, A.: The varifold representation of nonoriented shapes for diffeomorphic registration. SIAM Journal on Imaging Sciences **6**(4), 2547–2580 (2013)
7. Donohue, M.C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R.G., Raman, R., Gamst, A.C., Beckett, L.A., Jack Jr, C.R., Weiner, M.W., Dartigues, J.F., et al.: Estimating long-term multivariate progression from short-term data. Alzheimer's & Dementia **10**(5), S400–S410 (2014)
8. Duflo, M.: Algorithmes stochastiques. Springer Berlin (1996)
9. Durrleman, S., Prastawa, M., Charon, N., Korenberg, J.R., Joshi, S., Gerig, G., Trouvé, A.: Morphometry of anatomical shape complexes with dense deformations and sparse parameters. NeuroImage **101**, 35–49 (2014)
10. Jedynak, B.M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B.T., Raunig, D., Jedynak, C.P., Caffo, B., Prince, J.L., et al.: A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. Neuroimage **63**(3), 1478–1486 (2012)
11. Lorenzi, M., Ayache, N., Pennec, X.: Schilds ladder for the parallel transport of deformations in time series of images. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 463–474. Springer (2011)
12. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. Journal of mathematical imaging and vision **24**(2), 209–228 (2006)
13. Schiratti, J.B., Allassonnière, S., Colliot, O., Durrleman, S.: A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. The Journal of Machine Learning Research **18**(1), 4840–4872 (2017)

14. Singh, N., Hinkle, J., Joshi, S., Fletcher, P.T.: Hierarchical geodesic models in diffeomorphisms. International Journal of Computer Vision **117**(1), 70–92 (2016)
15. Vaillant, M., Glaunès, J.: Surface matching via currents. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 381–392. Springer (2005)