Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes

Pierre Fernique, Jean-Baptiste Durand, Yann Guédon

Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes

Pierre Fernique, *University of Montpellier 2, I3M and CIRAD, UMR AGAP and Inria, Virtual Plants*, pierre.fernique@inria.fr
Jean-Baptiste Durand, *University of Grenoble Alpes, Laboratoire Jean Kutzmann and Inria, Mistis*, jean-batiste.durand@imag.fr
Yann Guédon, *CIRAD, UMR AGAP and Inria, Virtual Plants*, guedon@cirad.fr

Abstract. We address the inference of discrete-state models for tree-structured data. Our aim is to introduce parametric multitype branching processes that can be efficiently estimated on the basis of data of limited size. Each generation distribution within this macroscopic model is modeled by a partially directed acyclic graphical model. The estimation of each graphical model relies on a greedy algorithm for graph selection. We present an algorithm for discrete graphical model which is applied on multivariate count data. The proposed modeling approach is illustrated on plant architecture datasets.

Keywords. Partially directed graphical model, graph selection, multivariate discrete distribution, tree pattern, branching process, plant architecture, multivariate count data

1 Introduction

We consider discrete-state stochastic processes indexed by a rooted tree. Our aim is to introduce parametric models that can be efficiently estimated on the basis of data of limited size and that are easily interpretable. These models rely on local dependency assumptions between parent and child vertices and belong to the family of multitype branching processes (MTBPs). In our practical setting of plant architecture analysis, the combinatorics induced by the variable and high number of child vertices in each state induces an inflation in the number of model parameters. We thus introduce parametric MTBPs incorporating parsimonious discrete graphical models for each generation distribution. In order to have interpretable results, we propose to focus on a family of multivariate discrete generation distributions such that:

- child states that tend to appear simultaneously or on the contrary to be incompatible can be identified,
- multivariate parametric distributions can be used since the direct estimation of probability masses on the basis of multivariate counts is unreliable except for very large data sets.
- these multivariate parametric distributions can have zero-inflated and right-skewed marginals, so that multivariate Gaussian distributions are not appropriate.
- these multivariate parametric distributions can be easily simulated and probability masses can be easily computed in order to investigate hypotheses on generation distributions and long range pattern formation in trees.

To achieve this goal, an approach based on probabilistic graphical models [8] to represent the conditional independence relationships for each generation distribution is considered. Three kinds of graphical models are usual: undirected (UG), directed acyclic (DAG), and partially directed acyclic graphical (PDAG) model.

Methods for graph identification were proposed for UGs, using either frequencies to directly estimate probability masses (so-called nonparametric estimation) or mutual information – see [10] and references therein. Under a multivariate Gaussian distribution assumption, an approach based on a L_1 penalization (Lasso) was proposed in [5], with some extension to Poisson distributions and more generally to GLMs [13].

Specific models and methods were developed for DAGs. Most methods for graph identification in DAGs are based on exploring the set of possible graphs using some heuristic (e.g. hill climbing [1]) and by scoring the visited graphs (e.g. using BIC), the graph with highest score being selected – see [8] for a review.

PDAGs, which generalize both UGs and DAGs, have been considered less often in the literature. In such models, both marginal independence relationships and cyclic dependencies between quadruplets of variables (at least) can be represented. A family of such models was proposed using conditional Gaussian distributions, but the problem of graph identification was not addressed [2]. We choose here to use discrete parametric PDAGs to model generation distribution in MTBPs and present a graph identification procedure for PDAGs.

2 Discrete PDAG modeling of generation distributions in MTBPs

Data of interest are tree-indexed sets $\boldsymbol{x} = (x_t)_{t\in\mathcal{T}}$ where $\mathcal{T} \subset \mathbb{N}$ is the set of vertices of a rooted tree graph $\tau = (\mathcal{T}, \mathcal{A})$ and $\mathcal{A} \subset \mathcal{T} \times \mathcal{T}$ the set of directed edges representing lineage relationships between vertices. By convention, the root of the tree graph has index 0. Let $x_t \in \mathcal{V} = \{0, \ldots, K-1\}$ denote the label of vertex t. Let $pa\,(.)$ denote the parent of a vertex, $ch\,(.)$ the children set of a vertex, $an\,(.)$ the ancestor set of a vertex and $de\,(.)$ the descendant set of a vertex. These notations also apply to set of vertices – see [8] for graph terminology. We here assume that x_t (resp. $\boldsymbol{x}$, τ) is the outcome of a discrete random variable X_t (resp. discrete random vector $\boldsymbol{X}$, random rooted tree T).

MTBPs are based on local dependency assumptions between parent and child vertices, more precisely on the following Markovian property – children are independent of their non-

descendants given their parent

$$\forall t \in \mathcal{T},\ \boldsymbol{X}_{ch(t)} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{T}\setminus de(t)} \Big| X_{pa(t)}\ ,$$

and a permutation invariance property – see [6] for details – in order to obtain a more parsimonious model. As a consequence, the joint distribution can be factorized as follows

$$P\left(T = (\mathcal{T}, \mathcal{A}), \boldsymbol{X} = \boldsymbol{x}\right) \propto P[X_0 = x_0] \prod_{t \in \mathcal{T}} P\left(\boldsymbol{N}_t = \boldsymbol{n}_t | X_t = x_t\right), \tag{1}$$

where $\boldsymbol{N}_t | X = x_t$ is the discrete random vector of the number of children of t in each state given x_t. Therefore the outcome to model is a discrete random vector $\boldsymbol{N}_t$ for each vertex

$$\boldsymbol{n}_t = \left(|\{s \in ch\left(t\right) | X_s = k\}|\right)_{k \in \mathcal{V}}.$$

MTBPs are thus specified by K discrete multivariate generation distributions.

We here propose an extension to PDAGs to model these generation distributions. This extension is based on an enlarged family of discrete parametric distributions incorporating multivariate generalizations of the classical univariate discrete parametric distributions: multinomial, negative multinomial and multivariate Poisson [7] distributions and corresponding regressions. Since we focus on a single generation distribution, we will omit in the following the tree indexing and parent state conditioning of each factor in (1). The class of considered PDAGs is such that the generation distribution factorizes as [9]

$$P(\boldsymbol{N} = \boldsymbol{n}) = \prod_{c \in \mathcal{C}} P(\boldsymbol{N}_c = \boldsymbol{n}_c | \boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)}), \tag{2}$$

where $\mathcal{C}$ denotes a partition of $\mathcal{V}$ such that in each subset, the induced subgraph – so-called chain component – is a connected undirected graph and each subset is connected – if connected – by directed edges.

Usually for each c in $\mathcal{C}$, $P(\boldsymbol{N}_c = \boldsymbol{n}_c | \boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)})$ can be factorize as a product of clique factors [9]. But in the case of multinomial, negative multinomial and multivariate Poisson distributions or regressions, each chain component is complete. PDAGs where chain components are not cliques could be introduced using the UG framework [13]. In such UGs, the graph is in fact a cyclic bidirected graph. This renders far more difficult and less reliable the exploration of long-range patterns in such models as many normalization constants have to be computed (one for each predictor value for a given clique). Therefore we chose to consider PDAGs such that chain components are complete.

Definition 2.1. *A clique directed acyclic graph (CDAG) is a PDAG such that:*

- *each chain component is a clique,*
- *each vertex of a clique has the same parent set,*
- *each parent set belongs to the power set of cliques.*

A probabilistic PDAG model is defined by a PDAG and a specification of the factors in (2). Our approach to identify such models relies on efficient methods for CDAG search and for variable selection in regression. Proposition 2.1 establishes a connection between probabilistic PDAG, CDAG and regression models.

Proposition 2.1.
A probabilistic PDAG model such that:

- *each source vertex of the graph is associated with some univariate distribution chosen among the binomial, negative binomial and Poisson distributions and mixtures of such distributions.*
- *each non-singleton source component of the graph is associated with some multivariate distribution chosen among diverse extensions of the multinomial distribution, the multivariate Poisson distribution and mixtures of such distributions,*
- *each component of the graph with at least one parent is associated with the corresponding families of univariate and multivariate regression models defined above in the case of source components,*

has the same distribution as a CDAG associated with the same parametric families such that for each edge in the CDAG that is not in the PDAG, the corresponding regression coefficient is null.

Proof. Let $G = (\mathcal{V}, \mathcal{E})$ be a PDAG and $\tilde{G} = \left(\mathcal{V}, \tilde{\mathcal{E}}\right)$ be a CDAG with $\tilde{\mathcal{E}} = \mathcal{E}' \cup \mathcal{E}''$ – where $\mathcal{E}' \cap \mathcal{E}'' = \emptyset$ – such that

$$\mathcal{E}' = \{(u, v) \in \mathcal{E} | (v, u) \in \mathcal{E}\} \tag{3}$$

and

$$\mathcal{E}'' = \left\{(s, t) \in \mathcal{V} \times \mathcal{V} \middle| \exists\, (u, v) \in ne\,(s) \times ne\,(t) \cap \mathcal{E} \setminus \mathcal{E}'\right\} \tag{4}$$

where $ne\,(.)$ is denoting the set of neighbors of a vertex. Because of equation (3), $\tilde{G}$ has the same chain components as G, since $\mathcal{E}'$ is the set of undirected edges in both $\mathcal{E}$ and $\tilde{\mathcal{E}}$. Equation (4) implies that the set of directed edges in G is included in $\tilde{\mathcal{E}}$: only edges from the neighbors of a parent of a child clique are added to every child clique vertices. As setting the regression coefficient to 0 does not change the conditional distribution, the two models are equivalent. □

As a consequence of proposition 2.1, given a CDAG and using ML estimators combined with Lasso type estimators [12] for parametric regressions, we select among all PDAGs sharing the same CDAG a sparse PDAG solution with the previously introduced parametric distributions. Therefore the PDAG estimation task is performed using a graph search within a CDAG space which has a cardinal a little bit higher than the DAG space one but far less important than the PDAG space one (see table 1). This graph search can be achieved as for previous algorithms presented in [8] for DAGs using hill climbing, greedy search, first ascent or simulated annealing algorithms. For defining such an algorithm, lemma 2.2 specify how DAG operators (add/remove/reverse directed edges) can be applied to each CDAG. Since the space search graph is not connected using these 3 operators – chain components remain unchanged – two operators specific to CDAGs have been added: chain merging and splitting:

- A pair (c, c') of chain components of $\mathcal{C}$ such that

  $$\left[pa\,(c) = pa\,(c') \setminus c\right] \wedge \left[ch\,(c) \setminus c' = ch\,(c')\right]$$

  will be merged in one chain component c'' which results from the removal of one chain component.

- A vertex from a chain component c can be removed and set to be a parent or a child of c resulting into the addition of one chain component.

Lemma 2.2. *Let $\mathcal{M}$ be a vertex set and let DAG($\mathcal{M}$) (resp. CDAG($\mathcal{M}$)) denote the set of DAGs (resp. CDAGs) with vertex set $\mathcal{M}$ and Part($\mathcal{M}$) denote the set of partitions of $\mathcal{M}$. There is a one-to-one mapping between CDAG($\mathcal{V}$) and $\{DAG(p)|\ p \in Part(\mathcal{V})\}$.*

Proof. For $\mathcal{G} \in \text{CDAG}(\mathcal{V})$, let $\mathcal{C}(\mathcal{G}) \in \text{Part}(\mathcal{V})$ be the set of chain components of $\mathcal{G}$. Let $\sigma(\mathcal{G})$ be the DAG with vertex set $\mathcal{C}(\mathcal{G})$ and such that (c, c') is an edge if there exists an edge from $a \in c$ to $b \in c'$ in $\mathcal{G}$. It is easily seen that σ is a bijection from $\text{CDAG}(\mathcal{V})$ to $\{\text{DAG}(p)|\ p \in \text{Part}(\mathcal{V})\}$ since every chain component c of $\mathcal{G} \in \text{CDAG}(\mathcal{V})$ is a clique, and since all vertices in c have the same parents. □

Proposition 2.2.
Let b_K (resp. a_K) be the number of labeled CDAGs (resp. DAGs) of K vertices. One have:

$$b_K = \sum_{k=1}^{K} \left\{ {K \atop k} \right\} a_k \tag{5}$$

where $\left\{ {K \atop k} \right\}$ denote the Stirling number of second kind.

Proof. Consider a set of K vertices. The Stirling number of second kind gives the number of ways of partitioning such vertex set into k non-empty cliques. For each of these partitions, a DAG can be defined (see lemma 2.2) and there are a_k such labeled DAGs. We then just need to consider that the number of cliques can vary from 1 to K for CDAGs of K vertices to prove proposition 2.2. □

a_K	b_K	c_K	K
1	1	1	1
3	4	4	2
25	34	50	3
543	715	1,688	4
29,281	35,381	142,624	5
3,781,503	4,258,357	28,903,216	6
1,138,779,265	1,222,487,933	13,663,125,680	7
783,702,329,343	816,625,721,787	14,762,428,500,992	8

Table 1: Number a_K of DAGs, b_K of CDAGs and c_K PDAGs [11] from 1 to 8 vertices (see proposition 2.2)

3 Characterizing the apple tree irregular bearing phenomenon using MTBP and CDAG models

Recently, statistical indices have been proposed to characterize alternation in flowering at whole plant scale with a yearly time step for different apple tree cultivars [4]. A correlation has

been highlighted between synchronicity of flowering within plants, alternation along axes, and alternation at whole plant scale. However, little is known about structural factors that may induce heterogeneity in the fates (vegetative or flowering) of sibling shoots, and thus improve regularity at whole plant scale despite alternation along axes. Considering the methodology described in [3], a tree structure (see fig. 1) was built from the apple tree dataset provided by E. Costes (UMR AGAP, AFEF Team, Inra, Montpellier, France) in order to illustrate the interest of MTBPs to investigate this phenomenon.

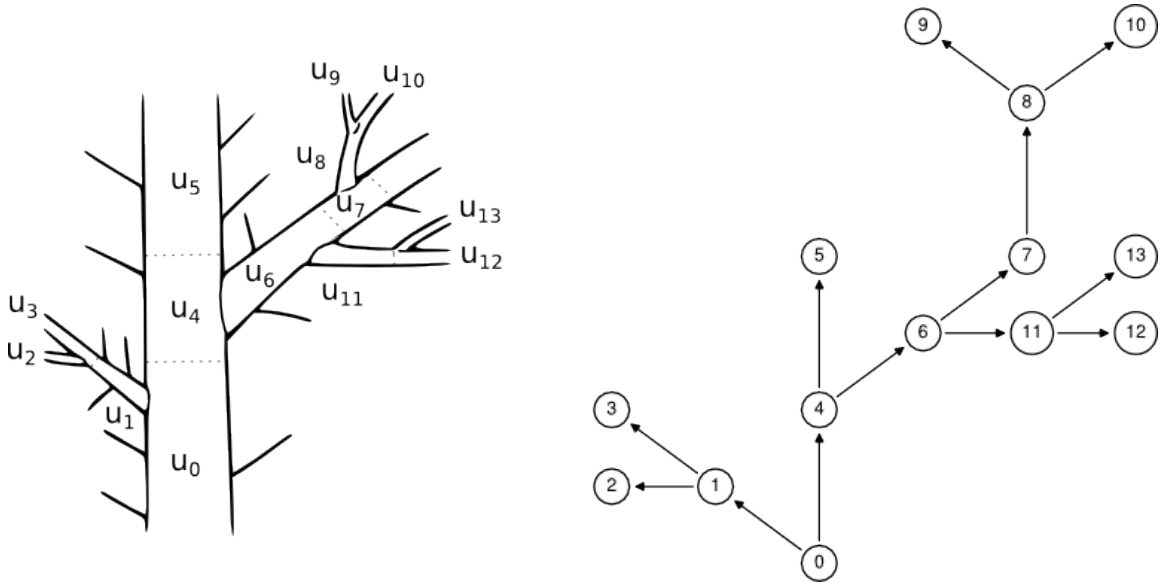

Figure 1: The tree is a formal representation of the plant topological information (drawing issued from [3]). Each label of this tree is the nature of the annual shoot.

MTBPs are used to model the number of flowering and vegetative shoots for parent shoots of different natures defined by their length and fate (see table 2). The aim is to identify parent states associated with homogeneous child fates from parents that may have heterogeneous child fates. As the dataset is composed of two trees per cultivar the objective is also to compare the two cultivars Fuji and Braeburn that have different behaviors regarding the irregular bearing phenomenon.

State	Length	Fate
0	Long	Vegetative
1	Long	Flowering
2	Medium	Vegetative
3	Medium	Flowering
4	Short	Vegetative
5	Short	Flowering

Table 2: Shoots state space and corresponding lengths and fates

CDAG-based generation distributions better fitted the data than DAG-based generation

distributions according to BIC. In the worst case, we obtained the same fit with CDAG-based and DAG-based generation distributions (see fig. 2).

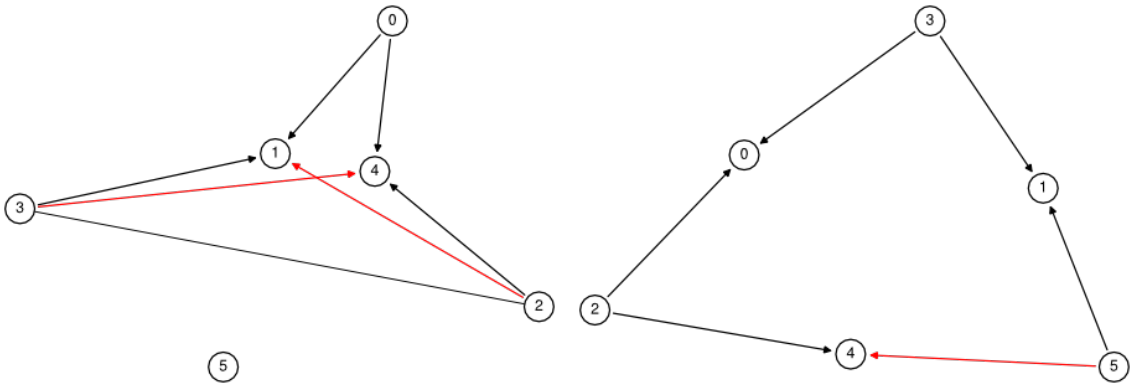

Figure 2: CDAG and DAG selected for the parent states 0 (left hand) and 1 (right hand) for the Braeburn cultivar. Edges associated with negative (resp. positive) covariances are in red (resp. black).

We obtained very contrasted graphs for the different parent states of a given cultivar. We also obtained different graphs for the two cultivars for some parent states. This was very informative for cultivar comparison (see fig. 2 and 3) The exam of the different graphs for a given cultivar highlights the more or less regular bearing behavior at the whole plant scale. Moreover comparing the graphs for the two cultivars leads to a better understanding of the biological functions underlying bearing behavior. This approach seems therefore promising to highlight pattern formation such as irregular bearing in tree structure development.

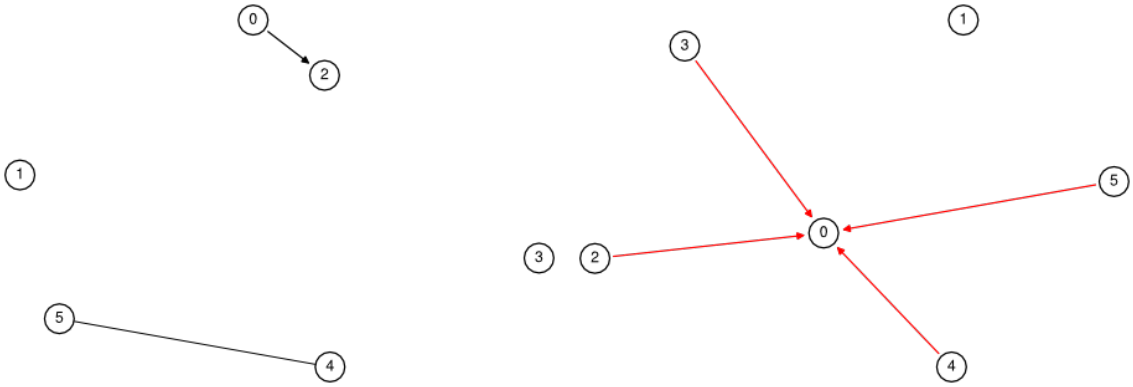

Figure 3: CDAG selected for the parent state 3 for the Braeburn (left hand) and the Fuji cultivar (right hand). Edges associated with negative (resp. positive) covariances are in red (resp. black).

Bibliography

[1] D.M. Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[2] Mathias Drton and Michael Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative markov property. *Scandinavian journal of statistics*, 33(2):247–257, 2006.

[3] J-B Durand, Y Guédon, Y Caraglio, and E Costes. Analysis of the plant architecture via tree–structured statistical models: the hidden Markov tree models. *New Phytologist*, 166(3):813–825, 2005.

[4] Jean-Baptiste Durand, Baptiste Guitton, Jean Peyhardi, Yan Holtz, Yann Guédon, Catherine Trottier, and Evelyne Costes. New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. *Journal of experimental botany*, 64(16):5099–5113, 2013.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[6] Patsy Haccou, Peter Jagers, and Vladimir A Vatutin. *Branching processes: variation, growth, and extinction of populations*. Cambridge University Press, 2005.

[7] D. Karlis. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.

[8] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[9] S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

[10] P.E. Meyer, F. Lafitte, and G. Bontempi. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC bioinformatics*, 9(1):461, 2008.

[11] Bertran Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, 270(1):267–278, 2003.

[12] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[13] Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1367–1375, 2012.