Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# On the characterization of flowering curves using Gaussian mixture models

Frédéric Proïa[a,*], Alix Pernet[b], Tatiana Thouroude[b], Gilles Michel[b], Jérémy Clotault[b]

[a] *Laboratoire Angevin de Recherche en Mathématiques – UMR 6093, Université d'Angers, Département de mathématiques, Faculté des Sciences, 2 Boulevard Lavoisier, 49045 Angers cedex, France*
[b] *IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France*

## HIGHLIGHTS

- A statistical methodology applied to the characterization of flowering curves using Gaussian mixture models is proposed.
- We suggest our own selection criterion to take into account the lack of symmetry of most of the flowering curves.
- A principal component analysis is conducted on a set of indicators derived from the statistical approach.
- Results suggest the lack of correlation between reblooming and flowering precocity.

## ARTICLE INFO

*Article history:*
Received 19 October 2015
Received in revised form
8 March 2016
Accepted 18 April 2016
Available online 22 April 2016

*Keywords:*
Flowering curves
Reblooming behaviour
Recurrent flowering
Gaussian mixture models
Longitudinal *k*-means algorithm
Principal component analysis
Characterization of curves
Classification of curves

## ABSTRACT

In this paper, we develop a statistical methodology applied to the characterization of flowering curves using Gaussian mixture models. Our study relies on a set of rosebushes flowering data, and Gaussian mixture models are mainly used to quantify the reblooming properties of each one. In this regard, we also suggest our own selection criterion to take into account the lack of symmetry of most of the flowering curves. Three classes are created on the basis of a principal component analysis conducted on a set of reblooming indicators, and a subclassification is made using a longitudinal *k*-means algorithm which also highlights the role played by the precocity of the flowering. In this way, we obtain an overview of the correlations between the features we decided to retain on each curve. In particular, results suggest the lack of correlation between reblooming and flowering precocity. The pertinent indicators obtained in this study will be a first step towards the comprehension of the environmental and genetic control of these biological processes.

## 1. Introduction and motivations

As it is explained by Putterill et al. (2004), matching the flowering period with the best climatic conditions is a crucial step for wild plants to obtain a high fertility rate. In agriculture, the amount of seeds and fruits produced by plants is directly related to their ability to produce a great number of flowers, hence flowering is extremely important for high-yields crops. For ornamental plants, obtaining a large number of flowers over the longest period of the year is an important breeding objective.

Plants present a large diversity of flowering patterns between taxa and suitable parameters are necessary to summarize these flowering profiles. Flowering curves, counting the number of flowers observed for a plant at regular time intervals, can be obtained from field scorings. Statistical methods have been rarely used to efficiently describe and compare flowering curves. As an example, regression curves have been used to fit flowering curves, but only for *once-flowering* plants whose curve shape skewed from normality (see Clark and Thompson, 2011). Especially in horticulture, annual flowering curves are sometimes much more complex. *Reblooming* – or *recurrent flowering* – plants are able to flower and fructify several times over the year. Such plants are found among several ornamental species, like irises, hydrangeas, daylilies or roses, but also in fruit-producing species like strawberry or raspberry plants.

For roses (*Rosa* sp. or genus *Rosa*), flowering traits are

* Corresponding author.
*E-mail addresses:* Frederic.Proia@univ-angers.fr (F. Proïa), Alix.Pernet@angers.inra.fr (A. Pernet), Tatiana.Thouroude@angers.inra.fr (T. Thouroude), Gilles.Michel@angers.inra.fr (G. Michel), Jeremy.Clotault@univ-angers.fr (J. Clotault).

http://dx.doi.org/10.1016/j.jtbi.2016.04.022
0022-5193/

particularly important, either for cut or garden roses. In Occident, the nineteenth century represents a golden age for rosebush breeding. It involved the creation of many cultivars, with the introduction of new traits in created hybrids, as explained by Oghina-Pavie (2015). Very early in this century, breeding activities have aimed at obtaining earlier – or later – flowering cultivars to increase the range of flowering periods (Oghina-Pavie, pers. comm.). Later in this century, the reblooming trait became the most important trait in rosebush breeding. Current modern roses result from crosses between reblooming Chinese roses and once-flowering European roses, obtained during the nineteenth century, according to Wylie (1954). By the number of created cultivars and by the diversification of flowering profiles, rosebush genetic resources of the nineteenth century are probably among the most interesting models for developing methodologies to characterize flowering curves.

The biological sample analysed in this article is composed of 329 exploitable flowering curves obtained in 2012 in the rose garden "Loubert" (Les Rosiers-sur-Loire, France). The studied genotypes were predominantly bred during the nineteenth century. For each genotype, the number of open flowers was counted almost each week between May 10 and November 15. For the most widespread case, were considered as open flowers developmental stages at or between these two following stages: (1) flower bud with at least one sepal detached from petals and at least one petal detached from the others (except for simple flowers, having five petals) and (2) flower with at least one petal which remains with original aspect and colour (see Fig. 1.1). The plant shape (sphere, cylinder or cone), circumference and height were measured for calculation of flower density (number of flowers per m$^2$). The mean number of flowers within an inflorescence was also counted. The dataset originally contained some irregularities and missing values, and different temporal lags between consecutive measures. All these issues have been carefully dealt with by the authors, but occasional presence of residual artificial values cannot be excluded.

In rosebush, the contemporary works of Iwata et al. (2012) and Kawamura et al. (2015) have highlighted the fact that flowering process is tightly linked to the branching process of the plant. In once-flowering cultivars, inflorescences are produced in the spring by the development of shoots from axillary buds of shoots from the previous year. Later in the year, new indeterminate shoots are produced and remain vegetative (having no flower). Inflorescence will develop the year after from axillary buds, issued of these vegetative shoots. In reblooming cultivars, either axillary buds will give inflorescence, or new determinate shoots terminated by an inflorescence will emerge successively from older shoots. Therefore, the best way to characterize rose flowering profile would be to differentiate the number of flowers produced by each shoot developed along the year. As an illustration of decomposition of the flowering shoot by shoot, Durand et al. (2013) tried to model biennial bearing in apple trees. For a large sample of elderly rosebushes with many shoots like the one studied in this article, this represents a huge and laborious work. Statistical methods are therefore needed to characterize flowering profiles (flowering date, flowering intensity, reblooming magnitude) from flowering curves obtained by counting flowers along the year in the whole plant. As for the characterization of reblooming, it is especially challenging to distinguish a long unique flowering period from several partially overlapping ones, corresponding to successive floral initiations.

Mixture models have for a long time been popular in life sciences, especially in biology and genetics, in fact since the seminal works of Pearson in the late nineteenth century. We guide the reader to the far from exhaustive mixtures applications to biology by Haley and Knott in 1992, Lynch and Walsh in 1998, Detilleux and Leroy in 2000, Boettcher et al. in 2005, Choi et al. in 2010, Shekofteh et al. in 2015, and references therein. The ability of a Gaussian mixture to split an apparently chaotic whole phenomenon into simple components and to highlight hidden structures was our main motivation to apply such models to flowering curves. Indeed, a rosebush is made of stems among which one may start to bloom while another starts to loose its flowers. On a whole set of branches, the resulting phenomenon is not suitably explained from a deterministic approach consisting in counting all variations from one week to the other. On the contrary, the *waves* mechanism of Gaussian mixture models seems to form a relevant alternative, as we will see in Appendix A. The paper is organized as follows. Sections 2 and 3 are devoted to the statistical tools that we intend to customize and to their applications on our dataset, respectively for the characterization and the classification of the flowering curves. In particular, a theoretical background is supplied, when necessary. Some concluding remarks are given in Section 4 and a schematic example is provided in Appendix A, to justify our choice of Gaussian mixture models.

## 2. An application of GMM to flowering curves

This section is devoted to the application of the *Gaussian Mixture Model* – shortened from now on *GMM* – on a set of flowering curves according to a statistical methodology that we will see in detail. Firstly, we need to supply a short theoretical background about GMM (see McLachlan and Basford, 1988; McLachlan and Peel, 2000 for more details).

### 2.1. The Gaussian mixture model

Consider a set $(X_1, \dots, X_n)$ of $n$ real-valued random variables that we want to divide in $k$ classes. For all $1 \le i \le n$, we denote by $Z_i$ the latent random variable in $\{1, \dots, k\}$ corresponding to the class of $X_i$. We suppose that $(Z_1, \dots, Z_n)$ are independent and have the same distribution as a random variable $Z$ such that, for all $1 \le j \le k$,

$$\mathbb{P}(Z = j) = \pi_j.$$

In addition, we suppose that for all $1 \le i \le n$, the random variable $X_i | \{Z_i = j\}$ has a $\mathcal{N}(\mu_j, \sigma_j^2)$ distribution and accordingly, $\pi_j$ stands for the proportion of the class $j$ in the whole population. For all $x \in \mathbb{R}$, the distribution of the mixture is

$$f_{\text{GM}}(x) = \sum_{j=1}^{k} \pi_j \, f(x|\mu_j, \sigma_j^2)$$

where $f(\cdot|\mu_j, \sigma_j^2)$ is the Gaussian distribution function with parameters $\mu_j$ and $\sigma_j^2$. If we consider that, given a subdivision in $k$ classes (that is, conditionally on the latent variables), the sequence $(X_1, \dots, X_n)$ is made of independent variables, then the (incomplete) log-likelihood for a set $\beta_k = (\pi_1, \dots, \pi_k, \mu_1, \dots, \mu_k, \sigma_1^2, \dots, \sigma_k^2)$ of $3k$ parameters is given for any observation $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ by

$$\ln \ell_{\text{GM}}(x|\beta_k) = \sum_{i=1}^{n} \ln\left( \sum_{j=1}^{k} \pi_j \, f(x_i|\mu_j, \sigma_j^2) \right) = \sum_{i=1}^{n} \ln f_{\text{GM}}(x_i). \tag{2.1}$$

The classic approach (see e.g. Day, 1969; Xu and Jordan, 1996) to estimate the $3k - 1$ parameters, considering the relation $\pi_1 + \dots + \pi_k = 1$, is to run the so-called *Expectation-Maximization algorithm* (Dempster et al., 1977) to maximize the above log-likelihood. The resulting estimator $\widetilde{\beta}_k$ is finally used to classify the observations via the Bayes' theorem. Namely,

$$\widetilde{\mathbb{P}}(Z_i = j | X_i = x_i) = \frac{\widetilde{\pi}_j \, f(x_i|\widetilde{\mu}_j, \widetilde{\sigma}_j^2)}{\sum_{\ell=1}^{k} \widetilde{\pi}_\ell \, f(x_i|\widetilde{\mu}_\ell, \widetilde{\sigma}_\ell^2)}$$