# Modelling photosynthetic parameters via mixed effects models and the Farquhar-van Cammerer-Berry model

Stephen D. Coleman

February 23, 2019

**Abstract**

The study of the photosynthetic process frequently involves analysis of net assimilation - intercellular $CO_2$ concentration (ACI) curves. These are used in the estimation of key parameters associated with the Farquhar-van Caemmerer-Berry (FvCB) [5] model:

- $V_{c_{max}}$: the rate of maximum Rubisco carboxylation;
- $J$: electron transport rate;
- $R_d$: daytime respiration; and
- $g_m$: mesophyll conductance.

Accurate, unbiased estimation of these parameters is a non-trivial exercise with the optimal method still a matter of debate within the field [6], [9], [7]. The problem of selecting suitable starting values for a non-linear model is thus complicated by the lack of unanimity on what is a "good" estimate of these. The core model we are using to estimate the parameters is based on Sharkey et al. [8].

The data collected for the purpose of modelling steady-state photosynthesis is inevitably repeated measurements on individual plants. This leads to correlation between measurements, a fact that violates the assumptions in the fixed effects models used in estimating the parameters of interest ([4], [2]). We introduce the use of non-linear mixed effects models in keeping with Qian et al. [7] to overcome this issue, and extend on their work by considering more modern versions of the FvCB model.

***Keywords***— ACI curves, parameter estimation, FvCB model, non-linear mixed effects models

## 1 Mixed-effects models

### 1.1 Linear mixed-effects models

Traditional parametric models incorporate only *fixed effects*. That is, they have a set of parameters describing with the entire population. For example, consider a system of $n \in \mathbb{N}$ observations of dependent variable, $Y = (y_1, \ldots, y_n)$, and the $n \times p$ matrix of associated independent variable measurements, $X = \{x_{i,j}\}$ for $i \in [1, n], j \in [1, p]$ for some $p \in \mathbb{N}$, and a $p$-vector of weights, $\beta$, represented:

$$Y = X\beta + \epsilon \tag{1}$$

Here $X$ is assumed to contain a column of 1's in the first position (hence the +1 in the description of its dimensionality) and $\epsilon$ is the $n$-vector of associated errors. We assume that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for

$i = (1, \ldots, n)$. If we consider the intuitive case of *biological* and *technical replicates*, where we have $n$ biological replicates and for the $i^{th}$ sample $m_i$ associated technical replicates. As each of the $m_i$ measurements is on the same sample we expect there to be a non-negligible correlation between the $m_i$ technical replicates for each $i$. The model in (1) does not allow for the within group effects due to the correlation between technical replicates. Consider the simplest possible model for the fixed effects model including only the intercept:

$$y_{ij} = \beta + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{2}$$

Some of this model's limitations become apparent if one considers the case of unbalanced data. Continuing the previous example of biological and technical replicates, consider the case that $m_i \neq m_j$ for any $i, j \in (1, \ldots, n)$. In this case the model is skewed by the within sample data rather than by the true observations, the biological replicates. A possible solution is the inclusion of a individual intercept for each group of technical replicates, accommodating the within-sample variability, or *random effects*:

$$y_{ij} = \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{3}$$

While this better describes the observed data, it has certain inherent flaws; most notably the number of parameters scales linearly with the number of observations and the model only describes the measurements included in the sample. Consider as a solution a combination of these models, containing both a sample mean, $\beta$, and a random variable for each group representing the deviation from the population mean, $b_i$, i.e. a *mixed effects* model:

$$y_{ij} = \beta + b_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{4}$$

The linear mixed effects model described in (4) contains information at both a population level (in the fixed effects) but also at the individual level (in the random effects).

For now we assume $b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2)$ for $i = 1, \ldots, n$. This means that the variance of the observations is divided into two parts, $\sigma_b^2$ for the biological variability and $\sigma^2$ for the technical variability:

$$b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2), \quad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{5}$$

The assumption of normality can be modified if deemed inappropriate and it is possible to generalise the model to allow for heteroscedasticity.

The $b_i$ are called *random effects* as they are associated with the experimental unit and selected at random from the population of interest (at least in theory, obviously there are limitations on this particularly in the area of medicine) They represent that the effect of choosing the sample $i$ is to shift the mean expression of $Y$ from $\beta$ to $\beta + b_i$ - i.e. they effect a deviation from an overall mean. Technical replicates share the same random effect $b_i$ and are correlated. The covariance between technical replicates on the same experimental unit is $\sigma_b^2$; this corresponds to a correlation of $\frac{\sigma_b^2}{(\sigma_b^2 + \sigma^2)}$.

## 1.2 Non-linear mixed effects models

Non-linear mixed effects models, also known as *non-linear hierarchical models*, are an extension to the more traditional linear mixed effects models. They are used in scenarios where all of the following features are present [3]:

1. Repeated observations of a continuous variable on each of several *experimental units* (in our case these are individuals, thus individual is considered equivalent to empircial unit in the following section) over time or another condition (e.g. measurements at given heights on a tree) (the *condition variable*);
2. We expect the relationship between the response variable and the condition variable to vary across individuals; and

3. Availability of a scientifically relevant model characterising the behaviour of the individual response in terms of meaningful parameters that vary across individuals and dictate variation in patterns of condition-response (for us this will be the Farquhar-van Cammerer-Berry model).

The final point from the list in 1.2 is where the non-linear aspect is introduced. This is often a mechanistic function describing a physical or chemical system (for example in toxicokineitcs physiologically-based pharmacokinetics models are used or HIV dynamics in "precision medicine"); that is the model is described by by meaningful, interpretable parameters rather than being an empirical best fit. It is expected that the mechanistic model will better describe data beyond the range of the measurements used here (whereas an empirical fit might describe the data over the range captured in the measurement data but then misbehave woefully beyond these boundaries).

The analysis tends to have the goal of understanding one or more of the following:

1. The "typical" behaviour of the phenomena (i.e. mean or median values) represented by the model parameters;
2. The variation of these parameters, and hence the phenomena, between individuals; and
3. If some of the variation is inherently associated with individual characteristics.

Individual level prediction can also be of interest (e.g. in medical treatment with highly individual reaction), but is less relevant to this project. From the individual level we are interested in investigating the level of variation between individuals and questioning if this is sufficiently small to allow the "all-purpose" models generally used in photosynthesis describing the parameters and systems of interest.

Consider an experiment involving repeated measurements of some response variable, $Y$, across a condition variable $T$ for $n$ individuals. Each individual's characteristics are recorded in $A$ (this is assumed to be time-independent measurements, like an individual's height over the course of a drugs trial) and possible additional initial conditions in $U$ (for example the dose of a drug the individual was receiving at time $t = 0$). Let $y_{i,j}$ denote the $j$th measurement of the response under condition $t_{i,j}$, $j = 1, \ldots, n_i$, for individual $i$, $i = 1, \ldots, n$ with additional initial conditions $u_i$. Thus:

$$\begin{array}{ll} Y & = (y_1, \ldots, y_n) \\ y_i & = \left(y_{i,1}, \ldots, y_{i,n_i}\right) \\ T & = (t_1, \ldots, t_n) \\ t_i & = \left(t_{i,1}, \ldots, t_{i,n_i}\right) \\ A & = (a_1, \ldots, a_n) \\ U & = (u_1, \ldots, u_n) \end{array} \tag{6}$$

We denote $x_{i,j} = \left(t_{i,j}, u_i\right)$. Often $T$ is time and $U = \emptyset$, but it might be the case that each $t_{i,j}$ is a $p_1$-vector of measurements and $u_i$ is a $p_2$-vector such that $p_1 + p_2 = p$, returning to the notation of (4). The assumption that the triplets $(y_i, u_i, a_i)$ are independent across $i$ is often included to reflect the belief that individuals are unrelated (this will hold for us, but might require more thought in other situations). For some function $f$ regulating the within-individual behaviour defined by a vector of parameters $\beta_i$ unique to individual $i$, we have:

$$y_{i,j} = f\left(x_{i,j}, \beta_i\right) + \epsilon_{i,j}, \quad j = 1, \ldots, n_i \tag{7}$$

For us this will be the Farquhar-van Cammerer-Berry model of steady-state photosynthesis. We assume that $\mathbb{E}\left(\epsilon_{i,j} | u_i, \beta_i\right) = 0$ and $\epsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for all $i, j$ (the assumption of normality can be relaxed, but this extension exceeds the reach of this project). This model in (7) is called the *individual level model*. To model the population parameters we consider $d$, a $p$-dimensional function depending on an $r$-vector of fixed parameters, or *fixed effects*, $\beta$, and a $k$-vector of *random effects*, $b_i$, associated with individual $i$:

$$\beta_i = d\left(a_i, \beta, b_i\right), \quad i = 1, \ldots, n \tag{8}$$

Here, the *population model* in (8) describes how $\beta_i$ varies among individuals due to both individual attributes $a_i$ and biological variation in $b_i$. We assume that the $b_i$ are independent of the $a_i$, i.e.:

$$\begin{aligned} \mathbb{E}(b_i|a_i) &= \mathbb{E}(b_i) = 0 \\ \mathbb{V}ar(b_i|a_i) &= \mathbb{V}ar(b_i) = D \end{aligned} \tag{9}$$

Here, $D$ is an unstructured covariance matrix and is common to all $i$. It characterises the degree of unexplained variation in the elements of $\beta_i$ and associations among them; the ubiquitous assumption is $b_i \sim \mathcal{N}(0, D)$. However, if this set of assumptions regarding the conditional distribution of $b_i$ on $a_i$ is found to be insufficient, then $b_i \sim \mathcal{N}\left(0, D(a_i)\right)$ is frequently used.

In (8), $\beta_i$ is considered to have an associated random effect, reflecting the belief that each component varies non-negligibly in the population even after systematic relationships with subject characteristics are accounted for. It may happen that "unexplained" variance in a component of $\beta_i$ may be very small in magnitude relative to that in the remaining elements. In this situation it is common to drop the negligible quantity entirely. This lacks biological sense as each parameter is part of the "scientifically relevant model" and thus is unlikely to have no associated unexplained variation. Hence, one must recognise that this omission of an element of $\beta_i$ is adopted to achieve numerical stability in fitting rather than to reflect belief in perfect biological consistency across individuals and analyses in the literature to determine whether elements of $\beta_i$ are fixed or random effects should be interpreted so.

## 1.3 Our model

Much of the preceding concepts and theory make for very pleasant reading, but we consider it useful to explicitly state our model and it's assumptions in one place.

First, recall the FvCB model:

$$A = \min\left\{A_c, A_j\right\} \tag{10}$$

$$= FvCB(C_c, \Gamma_*, I, K_c, K_o, O_{i,j}, J_{max}, V_{c_{max}}, \alpha, \theta, R_{d_i}) \tag{11}$$

Where $A_c$ and $A_j$ are two curves describing separate biochemical limits on the rate of photosynthesis:

$$A_c = V_{c_{max}} \frac{C_c - \Gamma_*}{C_c + K_c(1 + O/K_o)} - R_d \tag{12}$$

$$A_j = J \frac{C_c - \Gamma_*}{4C_c + 8\Gamma_*} - R_d \tag{13}$$

Furthermore:

$$J = \frac{\alpha \cdot I + J_{max} - \sqrt{\left(\alpha \cdot I + J_{max}\right)^2 - 4\theta \cdot \alpha \cdot I \cdot J_{max}}}{2\theta} \tag{14}$$

We plug this into (7) for each of our experimental effects (in Rachel Schipper's data, SIDE and TREATMENT) in line with Bates and Pinheiro [1]. Thus, we have:

- Our response variable, $Y$, is net photosynthesis rate, $A$;
- Our individual level function, $f(\cdot)$, is $FvCB(\cdot)$;
- Our measured condition variables, $X$, are $C_c, \Gamma_*, I, K_c, K_o$, and $O$; and
- Our fixed effects, $\beta$, are $J_{max}, V_{c_{max}}, \alpha, \theta$ and $R_d$.

Notice that $U = \emptyset$. Therefore our individual-level model is:

$$\begin{aligned} A_{i,j} &= FvCB(C_{c_{i,j}}, \Gamma_{*_{i,j}}, I_{i,j}, K_c, K_o, O_{i,j}, J_{max_i}, V_{c_{max_i}}, \alpha_i, \theta_i, R_{d_i}) + \epsilon_{i,j} \\ \epsilon_{i,j} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\ \mathbb{E}\left(\epsilon_{i,j} | J_{max_i}, V_{c_{max_i}}, \alpha_i, \theta_i, R_{d_i}\right) &= 0 \end{aligned} \tag{15}$$

For our specific version of (8), we cannot know in advance what kind of model we will use. This will be based on the ability of different models to converge. We expect, given the paucity of data that we will have a diagonal covariance structure with 0's in the non-diagonal entries imposed on our mixed effects model (as we expect that there is not enough data for more flexible formats to converge). This corresponds to a statement of independence between random effects which may not be the ideal assumption, and with this in mind we will attempt other formats. However we do know that our photosynthetic parameters of interest ($J_{max}$, $V_{c_{max}}$, $\alpha$, $\theta$ and $R_d$) will all be represented in this aspect of the model to consider each individual plant with unique photosynthetic parameters.

# References


[1] Douglas M Bates and Jose C Pinheiro. Computational Methods for Multilevel Modelling. page 30.

[2] Chandra Bellasio, David J Beerling, and Howard Griffiths. An Excel tool for deriving key photosynthetic parameters from combined gas exchange and chlorophyll fluorescence: Theory and practice: Descriptive modelling of gas exchange data. *Plant, Cell & Environment*, 39(6):1180–1197, June 2016. ISSN 01407791. doi: 10.1111/pce.12560.

[3] Marie Davidian and David M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387–419, December 2003. ISSN 1085-7117, 1537-2693. doi: 10.1198/1085711032697.

[4] Remko A. Duursma. Plantecophys - An R Package for Analysing and Modelling Leaf Gas Exchange Data. *PLOS ONE*, 10(11):e0143346, November 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0143346.

[5] G. D. Farquhar, S. von Caemmerer, and J. A. Berry. A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. *Planta*, 149(1):78–90, June 1980. ISSN 0032-0935, 1432-2048. doi: 10.1007/BF00386231.

[6] Dany P. Moualeu-Ngangue, Tsu-Wei Chen, and Hartmut Stützel. A new method to estimate photosynthetic parameters through net assimilation rate-intercellular space CO $_2$ concentration ( $A$ - $C_i$ ) curve and chlorophyll fluorescence measurements. *New Phytologist*, 213(3):1543–1554, February 2017. ISSN 0028646X. doi: 10.1111/nph.14260.

[7] T. Qian, A. Elings, J.A. Dieleman, G. Gort, and L.F.M. Marcelis. Estimation of photosynthesis parameters for a modified Farquhar–von Caemmerer–Berry model using simultaneous estimation method and nonlinear mixed effects model. *Environmental and Experimental Botany*, 82:66–73, October 2012. ISSN 00988472. doi: 10.1016/j.envexpbot.2012.03.014.

[8] Thomas D. Sharkey, Carl J. Bernacchi, Graham D. Farquhar, and Eric L. Singsaas. Fitting photosynthetic carbon dioxide response curves for C $_3$ leaves. *Plant, Cell & Environment*, 30(9):1035–1040, September 2007. ISSN 01407791, 13653040. doi: 10.1111/j.1365-3040.2007.01710.x.

[9] Xinyou Yin and Paul C. Struik. Theoretical reconsiderations when estimating the mesophyll conductance to CO $_2$ diffusion in leaves of C $_3$ plants by analysis of combined gas exchange and chlorophyll fluorescence measurements. *Plant, Cell & Environment*, 32(11):1513–1524, November 2009. ISSN 01407791, 13653040. doi: 10.1111/j.1365-3040.2009.02016.x.