Separate regression modelling of the Gaussian and Exponential components of an EMG response from respiratory physiology

Gianfranco Lovison[1], Christian Schindler[2]

[1] Department od Economics, Business and Statistics, University of Palermo, Italy
[2] Swiss Tropical and Public Health Institute, Basel, Switzerland; University of Basel, Basel, Switzerland

E-mail for correspondence: gianfranco.lovison@unipa.it

Abstract: If $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $Y_2 \sim Exp(\nu)$, with Y_1 independent of Y_2, then their sum $Y = Y_1 + Y_2$ follows an Exponentially Modified Gaussian (EMG) distribution. In many applications it is of interest to model the two components separately, in order to investigate their (possibly) different important predictors. We show how this can be done through a GAMLSS with EMG response, and apply this separate regression modelling strategy to a dataset on lung function variables from the SAPALDIA cohort study.

Keywords: Exponentially Modified Gaussian distribution; GAMLSS; Deconvolution.

1 Introduction

The sum of two independent r.v.'s, one Gaussian and one Exponential, follows an Exponentially Modified Gaussian (EMG) distribution. Such a distribution has found interesting applications in some specific areas: modelling inter-mitotic time in genetics (Golubev, 2009), response times in experimental psychology (Palmer et al., 2011), peaks in chromatography, but seems to have received very little attention in biostatistics. We show in this paper how to fit separate regression models to the two components of an EMG response through a GAMLSS, and apply this separate regression modelling strategy to one of the lung function variables which arise in spirometry.

2 The Exponentially Modified Gaussian distribution

If $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $Y_2 \sim Exp(\nu)$, where $\nu = E(Y_2)$, with Y_1 independent of Y_2, then their sum $Y = Y_1 + Y_2$ follows an **Exponentially Modified Gaussian** (EMG) distribution, and one can then write $Y \sim \mathcal{EMG}(\mu, \sigma, \nu)$.

By convolution, the p.d.f. of $Y \sim \mathcal{EMG}(\mu, \sigma, \nu)$ can be shown to be:

$$f_Y(y; \mu, \sigma, \nu) = \frac{1}{2\nu} \exp\left[\frac{1}{2\nu}(2\mu + \frac{\sigma^2}{\nu} - 2y)\right] \text{ erfc}\left(\frac{\mu + \frac{\sigma^2}{\nu} - y}{\sqrt{2}\sigma}\right) \quad (1)$$

where $\text{erfc}(z) = 1 - \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \exp(-t^2)dt$ is the complementary error function. Exploiting the known relation: $\text{erfc}(\frac{z}{\sqrt{2}}) = 2\Phi(-z)$, where $\Phi(\cdot)$ is the Standard Normal distribution function, (1) can be written in the following form, perhaps more familiar to statisticians:

$$f_Y(y; \mu, \sigma, \nu) = \frac{1}{\nu} \exp\left(\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right) \Phi\left(\frac{y - \mu}{\sigma} - \frac{\sigma}{\nu}\right) \quad (2)$$

This is the parameterisation used by the R library `gamlss` (Rigby and Stasinopoulos, 2007) and adopted in this paper. The following expressions for the first four moments can be easily derived:

$$E[Y] = \mu + \nu; \quad Var[Y] = \sigma^2 + \nu^2;$$

$$Sk[Y] = 2\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-\frac{3}{2}}; \quad Ku[Y] = 6\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-2}.$$

Our interest in the EMG distribution arose in the study of lung function variables, where it accommodates in a flexible way both the (positive) skewness and the "peakedness" which characterise such variables. This flexibility, along with the possibility of a mechanistic interpretation of its derivation as the convolution of a Gaussian and an Exponential distribution, have motivated our preference for this distribution over other well-fitting, but somewhat more complex and less interpretable, positively skewed distributions, in analysing the dataset presented in Sec. 4.

3 Regression models for the Gaussian and Exponential components of an EMG response

Suppose a response variable Y is known to be the sum of two unobservable components Y_1, Y_2, which are of substantive interest, and that two GLMs: $\mathcal{M}_1 : E[Y_1] = h_1(\boldsymbol{X}\boldsymbol{\beta}); Var[Y_1] = \phi_1 V(E[Y_1])$ and $\mathcal{M}_2 : E[Y_2] = h_2(\boldsymbol{Z}\boldsymbol{\gamma}); Var[Y_2] = \phi_2 V(E[Y_2])$ are set up for modelling the effects of explanatory variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ on the expected values of the two components; the model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be formed by the same, by partly different or by completely separated explanatory variables.
Clearly, in general, if only the "convoluted response variable" $Y = Y_1 + Y_2$ is available, there will be serious problems of identifiability and estimability of the parameters $(\boldsymbol{\beta}, \phi_1)$ and $(\boldsymbol{\gamma}, \phi_2)$ in the two separate GLMs, depending

on the degree of separation and orthogonality of $\boldsymbol{X}$ and $\boldsymbol{Z}$. This difficulty parallels the complexity of deconvolving the distribution of the sum of two r.v.'s.
From this point of view, an EMG response Y is a fortunate exception. As outlined above, the location parameter of the Gaussian component enters only in the expression of $E[Y]$, while, for fixed σ, the higher moments depend only on the location parameter ν of the Exponential component. This makes it possible to specify two separate regression models for the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, assuming σ unknown but fixed:

$$\boldsymbol{y} \sim EMG(\boldsymbol{\mu}, \sigma, \boldsymbol{\nu}) \quad (3)$$

$$\boldsymbol{\mu} = h_{\mu}(\boldsymbol{X\beta}) \quad (4)$$

$$\boldsymbol{\nu} = h_{\nu}(\boldsymbol{Z\gamma}) \quad (5)$$

and to consider (3), (4) and (5) as a GAMLSS with EMG response distribution (Rigby, Stasinopoulos, 2005).

4 Application to respiratory physiology

SAPALDIA (Swiss Cohort Study on Air Pollution and Lung and Heart Diseases In Adults) is a large population-based cohort study, initiated in 1991 in eight areas of Switzerland. Participants were between 18 and 60 years old at recruitment. They were re-examined in 2002 and 2010/11. Besides responding to a computer-based interview with detailed questions on respiratory health and allergies, lifestyle, socio-demographic characteristics, home and workplace environment, study participants also underwent several examinations, including lung function testing. Methodological details are provided in Martin et al. (1997). SAPALDIA spirometry data have been used to derive sex-, age- and height- based reference equations for lung function variables in adults (Brändli et al., 1996 and 2000). Since the focus of these analyses was on modeling percentile functions, quantile regression methods were applied. Later, with the advent of GAMLSS modelling and related software, it became possible to fit models with skewness and kurtosis parameters. The Global Lung Function Initiative used this new methodological framework to develop a global set of spirometric reference equations for adults and children taking into account differences according to geography and race (Cole et al., 2009, Quanjer et al., 2012).
Two fundamental outcome variables of spirometry (i.e., lung function testing) are FVC, the Forced Vital Capacity of the lung, and FEV_1, the Forced Expiratory Volume in the 1st second. We focus in this paper on the difference $FEV_{a1} = FVC - FEV_1$, where FEV_{a1} stands for "Forced Expiratory Volume after the 1st second".
An extensive exploratory analysis on FEV_{a1} has shown a surprisingly good fit of the EMG distribution to the observed data. It is not yet clear whether

this reflects a precise causal mechanism, related to the physiology of respiration. In any case, it is of interest to try to find out the determinants of the two components, the Gaussian and the Exponential, through the approach outlined in Sec. 3.
For the purpose of illustration, we fitted the GAMLSS defined in (3), (4) and (5) to the sub-sample of male non-smokers in the first (1991) SAPALDIA survey; in keeping with the default options in `gamlss`, we chose $h_\mu = identity$ and $h_\nu = log$. The results of the final model, chosen through a stepwise procedure based on AIC, are reported in Table 1. Inspection of the table shows that the individual characteristics (Age, Height and BMI) combine in different ways to determine the Gaussian and Exponential components. In particular, BMI has a strong, both linear and quadratic, effect on the Gaussian component, along with an interaction with Age, but no significant effect on the Exponential component.

TABLE 1. Parameter estimates for the EMG model

	Estimate	Std. Error	t value	p-value
Regression model for the Gaussian component				
Intercept	-13.2116	3.14420	-4.20	0.00002
Age	0.0503	0.00586	8.58	0.00000
Height	0.1118	0.03596	3.10	0.00189
BMI	0.1288	0.02031	6.34	0.00000
Age^2	-0.0001	0.00005	-3.42	0.00063
$Height^2$	-0.0002	0.00010	-2.69	0.00718
BMI^2	-0.0013	0.00042	-3.27	0.00105
Age:BMI	-0.0010	0.00024	-4.29	0.00001
$\log(\sigma)$	-1.231	0.02141		
Regression model for the Exponential component				
Intercept	-5.8710	0.82903	-7.08	0.00000
Age	-0.0368	0.01581	-2.33	0.01978
Height	0.0290	0.00430	6.74	0.00000
Age^2	0.0005	0.00019	2.80	0.00500

An insightful way of presenting this model is to plot the two estimated component densities for a subject with a given combination of explanatory variables. As an example, in Figure 1 the plots on the same row have the same combination of Age and Height (top row: Age=20 yrs., Height=175 cm.; bottom row: Age=60 yrs., Height= 195 cm.), and therefore they have the same Exponential component. The left and right plot in each row differ only by BMI (left panel: BMI=24 kg/m^2, right panel=48 kg/m^2), and therefore their comparison helps to visualise the role of BMI, which affects only the Gaussian component. From inspection of these plots, it is evident

how older and taller people have a "flatter" (i.e. with larger mean) Exponential component, and also a more marked effect of BMI on reducing the mean of the Gaussian component.

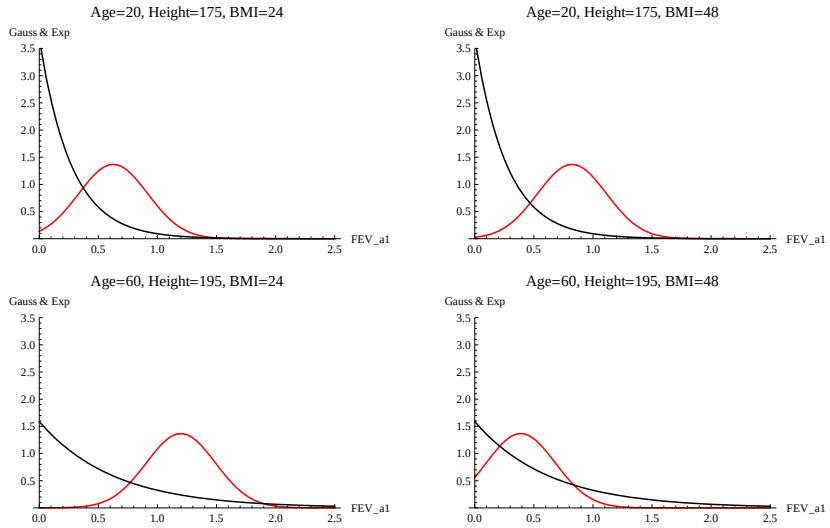

FIGURE 1. Estimated Gaussian and Exponential components for four exemplary individuals

The interplay of Age, Height and BMI in determining the two component distributions can be appreciated in Figure 2, where we report the estimated Gaussian and Exponential components for the two "extreme" individuals (i.e. having the two largest and smallest combinations of estimates $(\hat{\mu}, \hat{\nu})$) in our sample: in the left panel, a 51 years old man 197 cm. tall and with BMI = 27.1 kg/m^2; in the right panel, a 21 years old man, 164 cm. tall and with BMI = 19.3 kg/m^2.

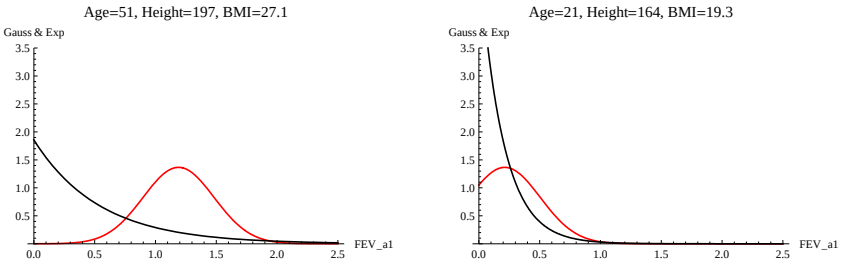

FIGURE 2. Estimated Gaussian and Exponential components for two "extreme" individuals

The combined effect of the three variables yields larger values of FEV_{a1} in

the older, taller and overweight subject in the left panel compared to the younger, shorter and normal weight subject in the right panel: this is the consequence of both the Exponential and the Gaussian components being shifted to the right for the latter compared to the former. In interpreting these findings, one should keep in mind that a large value of the $\frac{FEV_{a1}}{FEV_1}$ ratio is an indicator of obstructed expiration.

References

Brändli, O., Schindler, C., Künzli, N., et al. (1996). Lung function in healthy never smoking adults: reference values and lower limits of normal of a Swiss population. *Thorax*, **51**, 277 – 283.

Brändli, O., Schindler, C., Leuenberger, P., et al. (2000). Re-estimated equations for 5th percentiles of lung function variables. *Thorax*, **55**, 173 – 174.

Cole, T.J., Stanojevic, S., Stocks, J., et al. (2009) Age- and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in Medicine*, **28**, 880 – 898.

Golubev, A. (2009). Exponentially Modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. *Journal of Theoretical Biology*, **6**, 15 – 51.

Martin, B.W., Ackermann-Liebrich U., Leuenberger, P., et al. (1997). SAPALDIA: Methods and participation in the cross-sectional part of the Swiss Study on Air Pollution and Lung Diseases in Adults. *Sozial- und Präventivmedizin*, **42**, 67 – 84.

Palmer, E.M., Horowitz Todd, S., Torralba, A. et al. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology*, **37**, 58 – 71.

Quanjer, P.H., Stanojevic, S., Cole, T.J., et al., (2012) Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *European Respiratory Journal*, **40**, 1324 – 1343.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, **54**, 507 – 554.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.