This article was downloaded by: [Vlasios Voudouris]
On: 06 January 2012, At: 22:17
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

# Modelling skewness and kurtosis with the BCPE density in GAMLSS

Vlasios Voudouris [a b c], Robert Gilchrist [b], Robert Rigby [b], John Sedgwick [a] & Dimitrios Stasinopoulos [b]

[a] Centre for International Business and Sustainability, London Metropolitan Business School, London Metropolitan University, 84 Moorgate, London, EC2M 6SQ, UK

[b] Statistics, Operational Research and Mathematics Centre, London Metropolitan University, Holloway Road, London, N7 8DB, UK

[c] ABM Analytics, Suite 17 125, 145-157 St. John Street, London, EC1V 4PW, UK

Available online: 06 Jan 2012

To cite this article: Vlasios Voudouris, Robert Gilchrist, Robert Rigby, John Sedgwick & Dimitrios Stasinopoulos (2012): Modelling skewness and kurtosis with the BCPE density in GAMLSS, Journal of Applied Statistics, DOI:10.1080/02664763.2011.644530

To link to this article: http://dx.doi.org/10.1080/02664763.2011.644530

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings,

demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

*Journal of Applied Statistics*
2012, iFirst article

Taylor & Francis
Taylor & Francis Group

# Modelling skewness and kurtosis with the BCPE density in GAMLSS

Vlasios Voudouris[a,b,c]*, Robert Gilchrist[b], Robert Rigby[b], John Sedgwick[a] and Dimitrios Stasinopoulos[b]

[a]*Centre for International Business and Sustainability, London Metropolitan Business School, London Metropolitan University, 84 Moorgate, London EC2M 6SQ, UK;* [b]*Statistics, Operational Research and Mathematics Centre, London Metropolitan University, Holloway Road, London N7 8DB, UK;* [c]*ABM Analytics, Suite 17 125, 145-157 St. John Street, London EC1V 4PW, UK*

(*Received 4 June 2010; final version received 23 November 2011*)

This paper illustrates the power of modern statistical modelling in understanding processes characterised by data that are skewed and have heavy tails. Our particular substantive problem concerns film box-office revenues. We are able to show that traditional modelling techniques based on the Pareto–Levy–Mandelbrot distribution led to what is actually a poorly supported conclusion that these data have infinite variance. This in turn led to the dominant paradigm of the movie business that 'nobody knows anything' and hence that box-office revenues cannot be predicted. Using the Box–Cox power exponential distribution within the generalized additive models for location, scale and shape framework, we are able to model box-office revenues and develop probabilistic statements about revenues.

**Keywords:** GAMLSS; BCPE distribution; Pareto–Levy–Mandelbrot distribution; P-splines; model selection in GAMLSS

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

## 1. Introduction

Statistical modelling has developed greatly since Nelder and Wedderburn [15] introduced the unifying concept of the generalised linear model in the 1970s. Generalised additive models [11,27] advanced the concept by incorporating smooth functions of explanatory variables and, more recently, the simultaneous modelling of up to four distribution parameters (e.g. location, dispersion, skewness and kurtosis) has been advanced by Rigby and Stasinopoulos [22]. This paper benefits from the flexibility of the generalized additive models for location, scale and shape (GAMLSS) suite of software with its model selection facilities and diagnostic procedures that facilitate model selection.

*Corresponding author. Email: v.voudouris@londonmet.ac.uk

ISSN 0266-4763 print/ISSN 1360-0532 online
© 2012 Taylor & Francis
http://dx.doi.org/10.1080/02664763.2011.644530
http://www.tandfonline.com

Our particular data of interest are the cinema revenues of movies. These are highly skewed, with a small number of large revenue films coexisting alongside considerably greater numbers of smaller revenue films. Moreover, the skewed nature of these distributions appears to be an empirical regularity, with Pokorny and Sedgwick [17] dating this phenomenon back to at least the 1930s, making it an early example of a mass market long tail. Indeed, De Vany and Walls [5,6] commented on the consequential difficulty in modelling the dispersion, skewness and kurtosis of film revenues.

This difficulty is revisited and overcome here using the GAMLSS software and a unique micro data set of box-office revenues from the 1990s and 1930s. Our initial approach is to compare many competing models for the box-office revenues and specifically to compare these with the models based on the Pareto–Levy–Mandelbrot (PLM) distribution, which has dominated the modelling of box-office revenues since the work of De Vany and Walls [5].

The flexible four-parameter Box–Cox power exponential (BCPE) distribution, developed in [21], is used in Section 4.2 to provide a good fit of the 1990s' and 1930s' box-office revenues when explanatory variables are included in the model. The 1990s' data are more detailed, including data on the number of cinema screens and the size of the film distributor. Stepwise model selection is used to determine that the four parameters of the BCPE model can be modelled as smooth non-parametric functions of the opening box-office revenue and the number of screens. Automatic selection of the smoothing parameters (based on a penalised quasi-likelihood (PQL) estimation [2]) is used to allow smooth modelling of the location, scale, skewness and kurtosis parameters of the distribution of the response variable.

In summary, the distribution of box-office revenues is better approximated by the BCPE distribution than by the PLM distribution with its infinite variance. Moreover, the resulting smooth modelling of the distribution parameters as additive functions of the opening box-office revenues and the number of screens reveals a significant informational signal given the inflexible admission prices of the industry. Therefore, methodologically, the work presented here is consistent with the work of Haavelmo [9,10] and Mandelbrot [13] in the sense of a distribution that brings a degree of order and understanding at different time periods and/or scales of time.

Section 2 presents a discussion conducted on some of the more important statistical properties of the BCPE and PLM distributions. Section 3 discusses a model selection strategy for the distribution of the response variable, the modelling of the distribution parameters of the selected distribution and the optimal degree of smoothing for the additive smoothing functions. Section 4 begins with an account of the empirical revenue data of the 1990s and 1930s, which is followed by an assessment of how the flexibility of our semi-parametric model captures the long tails of the box-office revenues using a distribution with finite moments. Section 5 presents a summary of the results together with the conclusion that the BCPE distribution adequately fits film revenue data from different time periods and institutional arrangements.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

## 2. The GAMLSS framework: BCPE and PLM distributions

GAMLSS provides a very general and flexible way of modelling a (univariate) response variable as a function of explanatory variables. The distribution of the response variable can be any parametric distribution including highly skewed and kurtotic continuous and discrete distributions. The methodology is implemented in the free statistical software `R` [18]. The GAMLSS `R` implementation includes distributions with up to four parameters, denoted by $\mu$, $\sigma$, $\nu$ and $\tau$, which usually represent the location, scale, skewness and kurtosis shape parameters, respectively. All the parameters of the response variable distribution can be modelled using parametric and/or non-parametric smooth functions of explanatory variables, thus allowing modelling of the location, scale and shape parameters. We shall refer to $\mu_i, \sigma_i, \nu_i$ and $\tau_i$ as the *distribution parameters*. Rigby and Stasinopoulos [22] defined the original formulation of a GAMLSS model as follows.

For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating the distribution parameter $\theta_k$ to predictor $\eta_k$:

$$g_k(\theta_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{1}$$

where $\mathbf{X}_k$ and $\boldsymbol{\beta}_k$ are the matrices of the explanatory variables and parameters to be estimated, respectively, for $k = 1, 2, 3, 4$, and $h_{jk}$ are the smooth functions of explanatory variables $\mathbf{x}_{jk}$ for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$. The explanatory variables can be similar or different for each of the distributional parametera $\mu_i$, $\sigma_i$, $\nu_i$ and $\tau_i$. In this paper, we only use the P-splines of Eilers and Marx [8] as smooth functions $h_{jk}$, but the formulation in Equation (1) is general and any Gaussian Markov random field [23] formulation can be used here by setting the problem as a random effect model, that is, $h_{jk} = \mathbf{Z}_{jk}\gamma_{jk}$, where $\gamma_{jk} \sim N(0, \mathbf{Q}^{-1})$, with $\mathbf{Q}$ being a sparse precision matrix (see [22, Sections 3.2.1 and 3.2.3]).

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

### 2.1 *The BCPE distribution*

The BCPE distribution, BCPE$(\mu, \sigma, \nu, \tau)$, introduced by Rigby and Stasinopoulos [21] for centile estimation, provides a model for the response variable $Y$ exhibiting both skewness and kurtosis (leptokurtosis or platykurtosis).

The BCPE$(\mu, \sigma, \nu, \tau)$ distribution is specified for the positive random variable $Y$ through the transformed random variable $Z$ given by

$$Z = \begin{cases} \dfrac{1}{\sigma\nu}\left[\left(\dfrac{Y}{\mu}\right)^{\nu} - 1\right] & \text{if } \nu \neq 0, \\ \dfrac{1}{\sigma}\log\left(\dfrac{Y}{\mu}\right) & \text{if } \nu = 0, \end{cases} \tag{2}$$

for $0 < Y < \infty$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where the random variable $Z$ is assumed to follow a standard power exponential distribution with power parameter, $\tau > 0$, treated as a continuous parameter. [The parameterisation (2) was used by Cole and Green [4], who assumed a standard normal distribution for $Z$.]

The probability density function (pdf) of $Z$, a standard power exponential variable, is given by

$$f_Z(z) = \frac{\tau}{c 2^{(1+1/\tau)}\Gamma(1/\tau)} \exp\left\{-0.5\left|\frac{z}{c}\right|^{\tau}\right\} \tag{3}$$

for $-\infty < z < \infty$ and $\tau > 0$, where $c^2 = 2^{-2/\tau}\Gamma(1/\tau)[\Gamma(3/\tau)]^{-1}$. This parameterisation, used by Nelson [16], ensures that $Z$ has mean zero and standard deviation 1 for all $\tau > 0$. Note that $\tau = 1$ and $\tau = 2$ correspond to the Laplace (i.e. two-sided exponential) and normal distributions, respectively, while the uniform distribution is the limiting distribution as $\tau \to \infty$. (Strictly, the exact distribution of $Z$ in Equation (2) is a truncated standard power exponential distribution.) From Equation (2) the pdf of $Y$ is given by

$$f_Y(y) = f_Z(z)\left|\frac{\mathrm{d}z}{\mathrm{d}y}\right| = \frac{y^{\nu-1}}{\mu^{\nu}\sigma} f_Z(z) \tag{4}$$

for $y > 0$. The parameters $\mu$, $\sigma$, $\nu$ and $\tau$ may be interpreted as relating to location (median), scale (approximate coefficient of variation), skewness (transformation to symmetry) and kurtosis (power exponential parameter), respectively. Figure 1 shows how their values affect the curve of the BCPE distribution. Note that although $\mu$ can be interpreted as a measure of median box-office revenue, this is not shown because its effect on the BCPE distribution is a scaling along the $x$-axis.

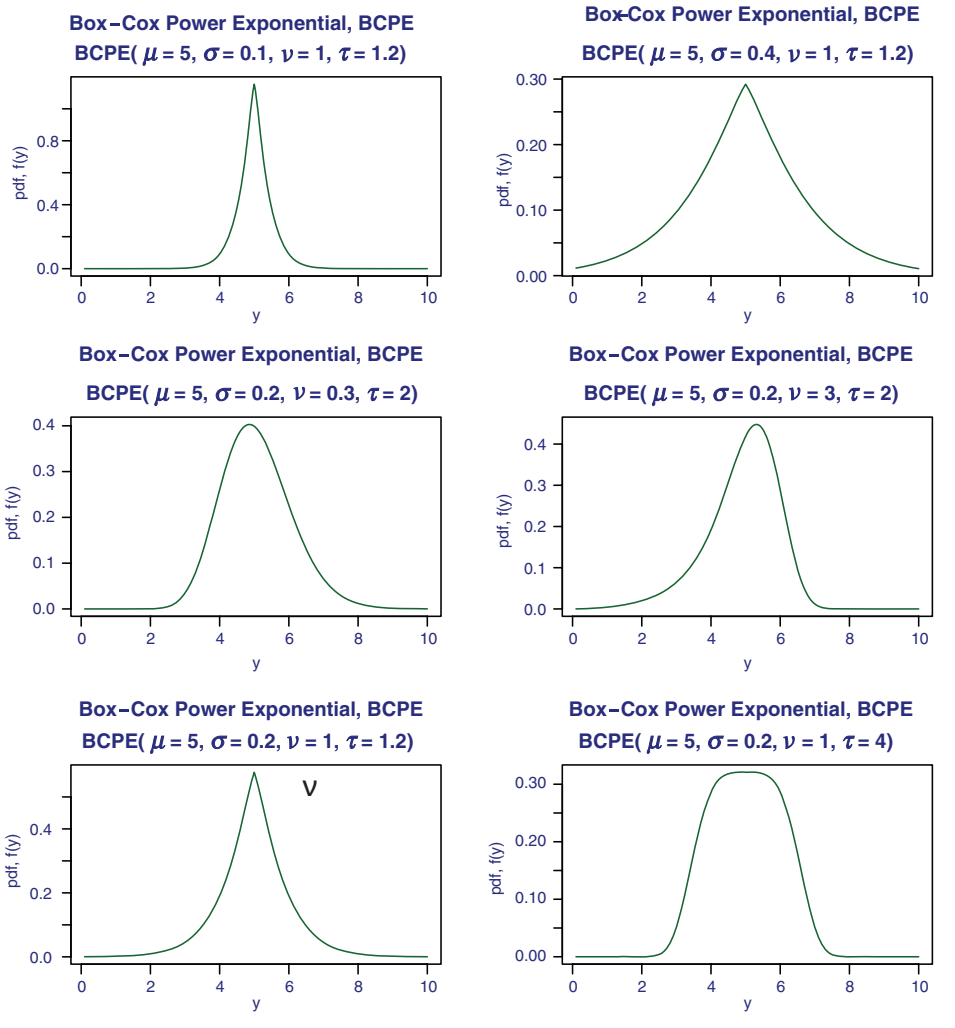


Figure 1. Examples of theoretical BCPE distributions.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

Figure 1 shows six plots of the BCPE distribution for different values of the parameters. All the six plots have the same $\mu$. The top two plots show the effect of changing $\sigma$ (while keeping the other three parameters fixed), that is, an increase in the spread of the distribution, the middle two plots show the effect of changing $\nu$, that is, a change in the skewness, and the bottom two plots show the effect of changing $\tau$, that is, a change in the kurtosis. Increasing $\mu$ increases the median of the distribution (by scaling of $Y$), since if $Y \sim \text{BCPE}(1, \sigma, \nu, \tau)$, then $\mu Y \sim \text{BCPE}(\mu, \sigma, \nu, \tau)$. Increasing $\sigma$ increases the spread (approximate coefficient of variation) of the distribution. Increasing $\nu$ decreases the skewness from positive skewness when $\nu < 1$ to symmetry when $\nu = 1$ and to negative skewness when $\nu > 1$. Increasing $\tau$ decreases the kurtosis. This is especially clear for the symmetric case when $\nu = 1$, with leptokurtosis when $\tau < 2$ (e.g. the Laplace, two-sided exponential when $\tau = 1$), mesokurtosis when $\tau = 2$ (the normal distribution) and platykurtosis when $\tau > 2$ (e.g. the uniform distribution limit as $\tau \to \infty$). Therefore, these four distribution parameters decide the shape of the curve: $\mu$ determines the 'signal', $\sigma$ determines the 'magnitude of the probabilities', $\nu$ determines the 'symmetry' and $\tau$ determines the 'fatness of the tails'. Note that $Y \sim \text{BCPE}(\mu, \sigma, \nu, \tau)$ implies that $Y = \mu\varepsilon$, where $\varepsilon \sim \text{BCPE}(1, \sigma, \nu, \tau)$ is a multiplicative

error model. Hence, the proposed model can be considered as an additive logarithmic model, $\log(Y) = \log(\mu) + \log(\varepsilon)$. Following [5], the multiplicative model suggests a strong form of interaction among filmgoers.

### 2.2 *The PLM distribution*

Pareto found that income was distributed according to a power law that was subsequently named after him. Levy showed that there is a class of distribution functions which follow the asymptotic form of the law of Pareto. This class of distributions was defined by Mandelbrot [14].

The pdfs of random variables $Y$ that follow the Pareto I or Pareto II distribution, respectively, are given by

$$\text{Pareto I} : f_Y\left(\frac{y}{\mu}, \sigma\right) = \frac{\mu^\sigma \sigma}{y^{\sigma+1}} \quad \text{for } y \geq \mu \text{ with } \mu > 0, \sigma > 0, \tag{5}$$

$$\text{Pareto II} : f_Y\left(\frac{y}{\mu}, \sigma\right) = \frac{\mu^\sigma \sigma}{(y+\mu)^{\sigma+1}} \quad \text{for } y \geq 0 \text{ with } \mu > 0, \sigma > 0. \tag{6}$$

It may be noted that the parameter $\sigma$ of the GAMLSS notation is usually denoted in the PLM literature by $\alpha$. We used $\sigma$ for consistency reasons in Section 2.1.

The Pareto I density function is zero for $y < \mu$, so $\mu$ provides a lower bound for $y$. The $r$th moment of $Y$ is given by $E(Y^r) = \mu^r(\sigma/(\sigma - r))$ for $\sigma > r$. The mean of the Pareto II distribution is given by $E(Y) = \mu/(\sigma - 1)$ for $\sigma > 1$, while the variance, skewness and kurtosis are the same as those for the Pareto I distribution.

The $\mu$ and $\sigma$ are parameters that do not represent the mean and standard deviation of $Y$. These parameters are, however, used to estimate the mean and variance of $Y$ when they exist. Asymptotically, as $y \to \infty$, the Pareto I and Pareto II distributions are equivalent.

The empirical findings ([5,6,25,26] and references therein) that $\sigma < 2$ mean that only the first moment exists for the Pareto I and Pareto II distributions. An implication of this is that the Pareto I distribution, which has dominated the modelling of box-office revenues in the literature, may not reveal the full structure and statistical properties of the box-office process that allows film rentals to rise and fall and booking periods to increase or contract.

## 3. Model strategy in GAMLSS

This section describes the model strategies adopted in this paper. Let $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$ represent a GAMLSS model as defined in Section 2. The components of $\mathcal{M}$ are defined as follows:

(i) $\mathcal{D}$ specifies the distribution of the response variable,
(ii) $\mathcal{G}$ specifies the set of link functions,
(iii) $\mathcal{T}$ specifies the terms appearing in all the predictors for $\mu$, $\sigma$, $\nu$ and $\tau$,
(iv) $\Lambda$ specifies the smoothing hyper-parameters which determine the amount of smoothing in the $h_{jk}()$ functions of Equation (1).

In the search for an appropriate GAMLSS model for any new data set, all the above four components have to be specified as objectively as possible.

We next discuss how the components $\mathcal{D}$, $\Lambda$, $\mathcal{T}$ and $\Lambda$ are specified in our analysis of the film data.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

### 3.1 *Component $\mathcal{D}$: selection of the distribution*

The selection of the appropriate distribution is done in two stages, the *fitting* stage and the *diagnostic* stage. The fitting stage involves the comparison of different fitted models using a generalised Akaike information criterion (GAIC). The GAIC is defined as $\text{GAIC}(k) = -2 \times \hat{\ell} + k \times \text{d}f$, where $\ell$ is the log-likelihood function and d$f$ are the effective degrees of freedom (edf) of the fitted model, respectively, and $k$ is a constant. We refer to $-2 \times \ell$ as the *global deviance*. The model with the smallest value of the criterion GAIC($k$) is then selected. The Akaike information criterion (AIC) [1] and the Schwarz Bayesian criterion (SBC) [24] are special cases of the GAIC($k$) corresponding to $k = 2$ and $k = \log(n)$, respectively. The two criteria, AIC and SBC, are asymptotically justified as predicting the degree of fit in a new data set, that is, approximations to the average predictive error. Justification for the use of SBC comes also as a crude approximation to Bayes factors [19,20]. In practice, it is usually found that while the original AIC is very generous in model selection, the SBC is more restrictive.

The diagnostic stage involves the use of *worm plots*. Worm plots were introduced by van Buuren and Fredriks [3] and are in effect detrended normal QQ plots of the quantile residuals (i.e. $z$-scores). The worm plot allows the detection of inadequacies in the model globally or within a specific range of an explanatory variable. van Buuren and Fredriks [3] proposed fitting cubic models to each of the detrended QQ plots, with the resulting constant, linear, quadratic and cubic coefficients, $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$ and $\hat{b}_3$, respectively, indicating differences between the empirical and model residual mean, variance, skewness and kurtosis, respectively, within the range in the QQ plot. They summarised their interpretations in their Table II. For model diagnosis, they categorised the absolute values of $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$ and $\hat{b}_3$ in excess of threshold values, 0.10, 0.10, 0.05 and 0.03, respectively, as misfits.

The normalised quantile residuals are defined as follows: given that the distribution of the response variable $f_Y(y;\boldsymbol{\theta})$ is fitted to observations $y_i$ for $i = 1, 2, \ldots, n$, the fitted normalised quantile residuals [7] are given by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}$ is the inverse cumulative distribution function of a standard normal variate and $\hat{u} = F_Y(y|\hat{\boldsymbol{\theta}})$ is the fitted cumulative distribution function, respectively. The advantage of normalised quantile residuals is that their true values $r_i$, $i = 1, 2, \ldots, n$, always have a standard normal distribution if the model is correct. (For discrete distributions, a randomisation of the quantile residuals is involved.)

### 3.2 *Component $\mathcal{G}$: selection of the link functions*

The selection of the link function is usually determined by the range of parameters in hand. For example, in the Pareto II distribution, both $\mu$ and $\sigma$ take values in the positive line, so a log link function seems a natural way of ensuring that they remain positive. For the BCPE distribution, we used log for $\mu$, $\sigma$ and $\tau$ and identity for $\nu$.

### 3.3 *Component $\mathcal{T}$: selection of the terms in the model*

Let $\mathcal{X}$ be the selection of all terms available for consideration. $\mathcal{X}$ could contain both linear and smoothing terms. For example, let $f_1$ and $f_2$ represent factors and $x_1$, $x_2$, $x_3$ and $x_4$ represent continuous explanatory variables, respectively. Then, $\mathcal{X} = \{f_1 * f_2 + s(x_1) + s(x_2) + s(x_3) + x_4\}$ allows second-order interactions for the factors and smooth functions for $x_1, x_2$ and $x_3$ and a linear term for $x_4$.

For a given distribution for the response variable, the selection of the terms for all the parameters of the distributions is done using a stepwise GAIC procedure. There are many different strategies that could be applied for the selection of the terms used to model the four parameters $\mu$, $\sigma$, $\nu$ and $\tau$. We now describe the two approaches that we employed in our analysis.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

#### 3.3.1 *Strategy A*

For all variables in $\mathcal{X}$ and for fixed distribution:

(1) use a backward GAIC selection procedure to select an appropriate model for $\mu$, with $\sigma$, $\nu$ and $\tau$ fitted as constants;
(2) use a forward selection procedure to select an appropriate model for $\sigma$, given the model for $\mu$ obtained in (1) and for $\nu$ and $\tau$ fitted as constants;
(3) use a forward selection procedure to select an appropriate model for $\nu$, given the models for $\mu$ and $\sigma$ obtained in (1) and (2), respectively, with $\tau$ fitted as constants;
(4) use a forward selection procedure to select an appropriate model for $\tau$, given the models for $\mu$, $\sigma$ and $\nu$ obtained in (1)–(3), respectively;
(5) use a backward selection procedure to select an appropriate model for $\nu$, given the models for $\mu$, $\sigma$ and $\tau$ obtained in (1), (2) and (4), respectively;
(6) use a backward selection procedure to select an appropriate model for $\sigma$, given the models for $\mu$, $\nu$ and $\tau$ obtained in (1), (5) and (4), respectively;
(7) use a backward selection procedure to select an appropriate model for $\mu$, given the models for $\sigma$, $\nu$ and $\tau$ obtained in (6), (5) and (4), respectively.

The final model may contain different subsets from $\mathcal{X}$ (i.e. not necessarily the same terms) for $\mu$, $\sigma$, $\nu$ and $\tau$.

#### 3.3.2 *Strategy B*

This strategy forces all the distribution parameters to have the same term. That is, a term from $\mathcal{X}$ is selected if its inclusion to the predictor in all the parameters improves the GAIC. The inclusion can be checked using forward, backward or stepwise procedure.

### 3.4 *Component* $\boldsymbol{\Lambda}$*: selection of the smoothing parameters*

Each smoothing term selected for any of the parameters of the distribution has one smoothing (or hyper) parameter $\lambda$ associated with it. We denote by $\Lambda$ the set of all smoothing parameters, for example, $\Lambda = \{\lambda_{\mu,1}, \lambda_{\mu,2}, \lambda_{\sigma,2}, \lambda_{\sigma,3}, \lambda_{\nu,1}\}$. Here, we assume that there are two terms in $\mu$ which need smoothing, $x_1$ and $x_2$, two terms in $\sigma$, $x_2$ and $x_3$, and one term in $\nu$, $x_1$.

The smoothing parameters can be fixed or estimated from the data. The standard way of fixing the smoothing parameters, as suggested in [11], is by fixing the edf for smoothing. The following are three of the different methods of estimating the smoothing parameters:

1. using generalised cross validation [27],
2. using GAIC [22],
3. using local maximum-likelihood method or a PQL method, as in [12].

The PQL method is the method used in our analysis. The method is implemented in the `R` function `pb()`, which uses a P-spline smoother [8]. Hence, the model terms were selected using the SBC, while the smoothing parameters (and hence corresponding edf) were chosen using PQL.

## 4. Empirical analysis

This section presents an empirical analysis of the box-office revenues with a number of distributions selected from the `gamlss.dist` package in order to develop and compare many

competing models for the box-office revenues and, specifically, to compare these with the models based on the Pareto I and Pareto II distributions.

The key aim here is to find a distribution that fits well two data sets representing different time periods and institutional arrangements (with the aim of searching for an intertemporal probabilistic law without '*ad hoc*' fixes). Because the 1990s' data set exhibits higher degree of kurtosis and skewness, we first use the 1990s' data set and then we test if the selected model also fits well the 1930s' data set. Using the `stepGAIC.A()` function in GAMLSS, we compare the models that best fit the 1990s' post-opening box-office revenue, conditional on the available explanatory variables such as the opening box-office revenue, number of screens, type of distributor and year of movie release.

### 4.1 *The 1990s' and 1930s' data sets*

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

The objective of the distribution system of movies is to maximise revenues. The institutional basis for doing this was different during the two periods. In the 1930s, films were first released to a small number of box-office rich *showcase* cinemas in metropolitan centres, where they built reputation, before being put out through time and place to a myriad of cinemas in particular localities, demarcated into runs based on their box-office capability. In effect, audiences in metropolitan centres expressed a time preference for movies, sometimes paying a premium for the privilege of an earlier screening, rather than waiting until the film appeared at a later date at lower status, generally less well-accoutred cinema in their locality.

Using weekly box-office data that were listed in the trade journal Variety, the 1930s' data set charts the diffusion of films among a sample of 104 first-run cinemas located in 24 cities across North America that were released between October 1934 and October 1936. These cinemas were top of the range, representing the first tier in the diffusion process described above. Altogether, the data set comprises the exhibition records of 969 films that received between them 11,016 screenings at these 104 cinemas. All revenues were converted into 1929 US dollars. Within the sample section, the most popular films opened and had extended runs in the very large box-office rich cinemas of New York and Chicago, before being released weeks later to the cinemas of the large provincial city centres such as Philadelphia and Los Angeles, after which they were screened weeks later in regional cities such as Tacoma and Denver. All of this occurred before wider distribution to second-run, third-run, fourth-run and so on cinemas everywhere.

By way of contrast, in the North American market during the 1990s, films were released simultaneously to as many screens as distributors and exhibitors thought desirable. The data set is derived from industry standard data sourced by Nielsen EDI for the North American market annually for 13 years from 1988 to 1999. All revenues are expressed in 1987 US dollars. During this period, 4164 films were released with revenues ranging from $145 to over $413 million, opening on between 2 and 3342 screens.

Figure 2(a) and (b) shows a box plot and histogram of the total box-office revenue, respectively, while Figure 2(c) plots the post-opening box-office revenue against the opening box-office revenue and Figure 2(d) plots the log of post-opening box-office revenue against the log of opening box-office revenue. Note that the total box-office revenue is equal to the post-opening box-office revenue plus the post-opening box-office revenue. The later plots try to explore the question whether the post-opening box-office revenue can be predicted given the information obtained on the first week of the box-office.

Figure 3, for example, shows descriptive plots of the box-office revenues in the 1930s, which shows that US theatrical movie market of the 1930s is also skewed and kurtotic (although to a lesser degree). The lower skewness and kurtosis should be reflected in the fitted model.

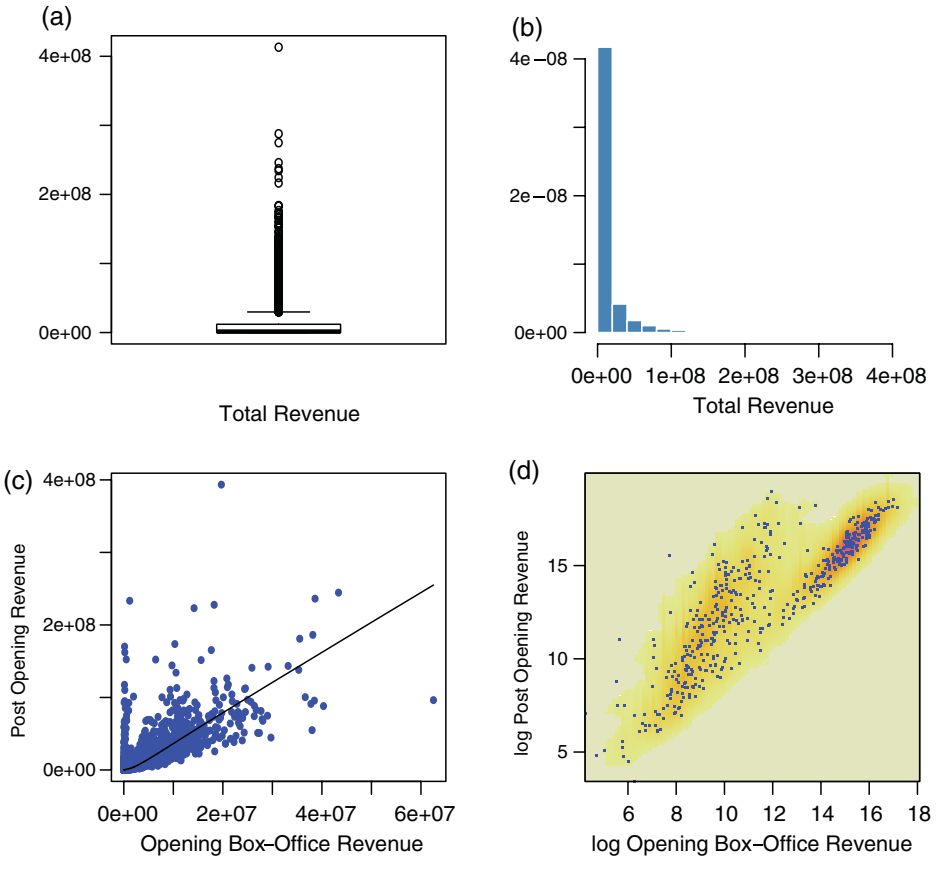


Figure 2. The 1990s' data: (a) and (b) total revenue and (c) and (d) post-opening revenues against opening revenue.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

### 4.2 *Modelling of distribution parameters*

Given that a set of plausible distributions have been identified from the marginal analysis, the parameters of the selected distributions are modelled as regression models. In particular, the distribution parameters are modelled using the penalised B-spline function as a way of understanding how the location, scale, skewness and kurtosis parameters of the distribution of the post-opening box-office revenues are affected by the opening box-office revenues, number of screens, distributor type (major or independent) and year of release. Unlike the conclusion drawn based on the Pareto distribution, the GAMLSS model based on the BCPE distribution suggests that we can, to a substantial extent, separate the high-return films from the low-return films early in the lifetime of a film.

Figure 4 shows the worm plot [3] of the BCPE distribution model. The BCPE distribution model fits very well the data. The lack of quadratic and cubic shape of the residuals indicates that the empirical skewness and kurtosis are appropriately captured by the BCPE distribution model.

Based on the model selection strategy A discussed in Section 3, the empirical GAMLSS-based model is $Y \sim \text{BCPE}(\mu, \sigma, \nu, \tau)$, where $Y = \log(\text{post-opening box-office revenue})$ and

$$\log(\mu) = pb(x_1, \text{d}f = 10) + pb(x_2, \text{d}f = 7) + x_3,$$
$$\log(\sigma) = pb(x_1, \text{d}f = 5) + pb(x_2, \text{d}f = 4) + x_4,$$

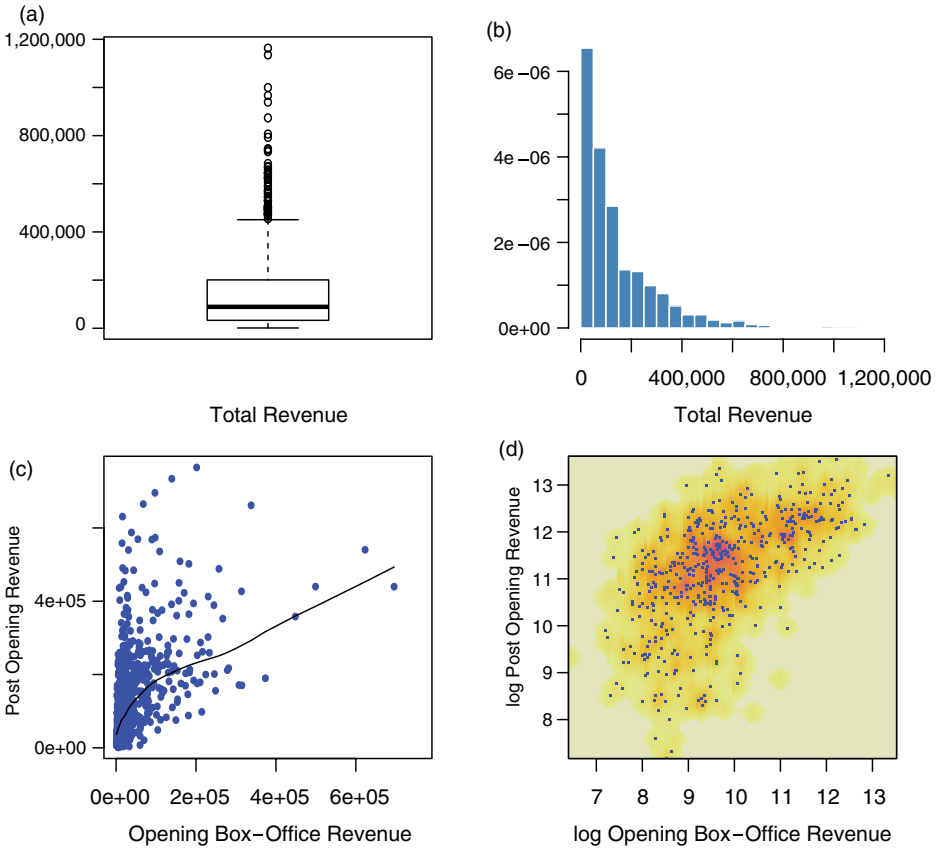


Figure 3. The 1930s' data: (a) and (b) total revenue and (c) and (d) post-opening revenues against opening revenue.

$$\nu = x_1 + pb(x_2, \mathrm{d}f = 2) + x_3, \tag{7}$$
$$\log(\tau) = 0.5705,$$

where $x_1 = \log$(opening box-office revenues), $x_2 = \log$(number of screens), $x_3 =$ type of distributor (major or independent) and $x_4 =$ year of movie release. For example, $pb(x_1, \mathrm{d}f = 10.5)$ is a non-parametric smoothing function of $x_1$ with an additional degree of freedom for smoothing on top of the linear term. The edf $= \mathrm{d}f + 2 = 12.5$, where the additional two degrees of freedom account for the linear term. When edf $\geq 3$, non-ergodic dynamics are in operation in the motion picture industry. It is important to note here that the regression models for the distribution parameters are different. This flexibility is important as the $\sigma$ (approximate coefficient of variation) of the BCPE distribution is not affected by the type of distributor (major or independent). This is consistent with subject-specific expertise.

Figure 5 graphically shows the dynamics of $\mu$, $\sigma$ and $\nu$ distribution parameters of Equation (7) based on the selected explanatory variables (lboopen = log of opening box-office revenues, lnosc = log of number of screens, type of distributor and year of release). Note that because $\tau = 1.77$, the distribution of the revenues is leptokurtotic.

The top three graphics in Figure 5 show that the partials for the $\mu$ (median) distribution parameter. We can observe that the higher the opening box-office, the higher the post-opening box-office. This effectively confirms the argument of the film producers about the opening power, despite the

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

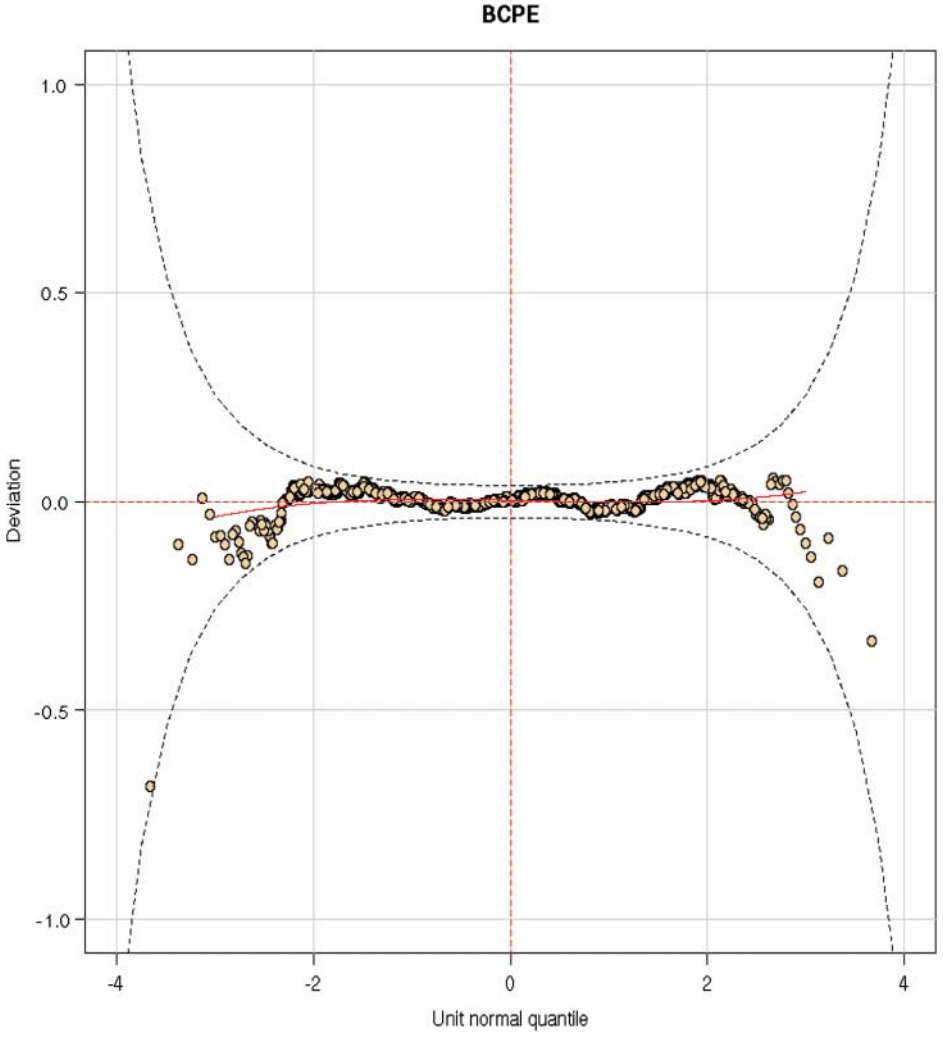


Figure 4. Worm plot of the BCPE distribution model.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

fact that this argument was refuted in academic literature based on the Paretian hypothesis. Given the opening box-office revenues, movies that opened with a large number of screens are less successful as shown by the middle graphic in the top row. Given the opening box-office revenues and the number of screens, major distributors also achieve higher post-opening box-office revenues.

The middle three graphics in Figure 5 show the partials for $\sigma$ (approximate coefficient of variation). The $\sigma$ of the BCPE distribution might be interpreted as a measure of the variation in a consensus opinion about a film. We can observe that the higher the opening box-office revenues, the greater the reduction in the coefficient (the decline is steeper in the broad middle of the opening box-office performance). Thus, the higher the opening box-office revenues, the lower the variation in a consensus option about a film. Given the opening box-office revenues, the coefficient of variation reduces when the number of opening screens is in the top half of the distribution of the number of screens. We can also observe that the coefficient of variation reduces with the passing of time, while the type of distributors has no significant impact. This suggests that both major

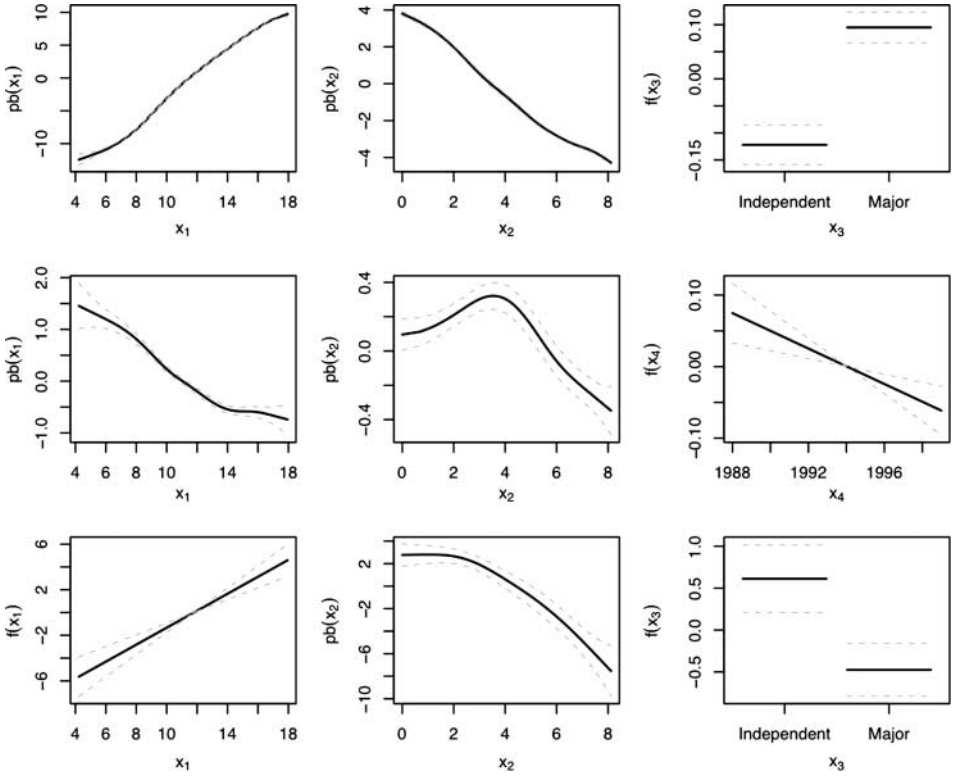


Figure 5. Dynamics of the BCPE distribution parameters of the fitted model.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

and independent film distributors collectively learn the best distribution strategies that reduce the 'risk' of a very rapid death process for a film.

The bottom three graphics in Figure 5 show the partials for the $\nu$ (transformation to symmetry). We can observe that as the number of opening screens increases, the skewness changes from negative to positive. Given the number of opening screens, the opening box-office revenues have the opposite effect. Finally, given the opening number of screens and the opening box-office revenues, independent distributors achieve higher degree of symmetry compared with the major distributors.

Figure 6 shows the fitted pdfs of 1990s' post-opening box-office revenues of the 1997 film 'Titanic' (the best performing film) and the 1998 film 'Touch Me' (a relatively poor performing film). Thus, by using GAMLSS, the shape of distribution of the post-opening box-office revenue can be estimated early in the lifetime of a film. The substantive implication of applying the GAMLSS model using the BCPE distribution is that probabilistic statements can be used (a) to correctly price future contracts indexed on the films' performance and (b) to adjust dynamically post-opening diffusion arrangements.

As one aim of this paper is to find a flexible distribution for the box-office revenues, we fitted the BCPE distribution to the box-office revenues of the 1930s' market, and the diagnostic plot shows an adequate fit across the data set. Figure 7 shows the fitted model for the selected 'opening box-office revenues' superimposed on a smoothed scatterplot diagram as a way of demonstrating the flexibility of the distribution.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

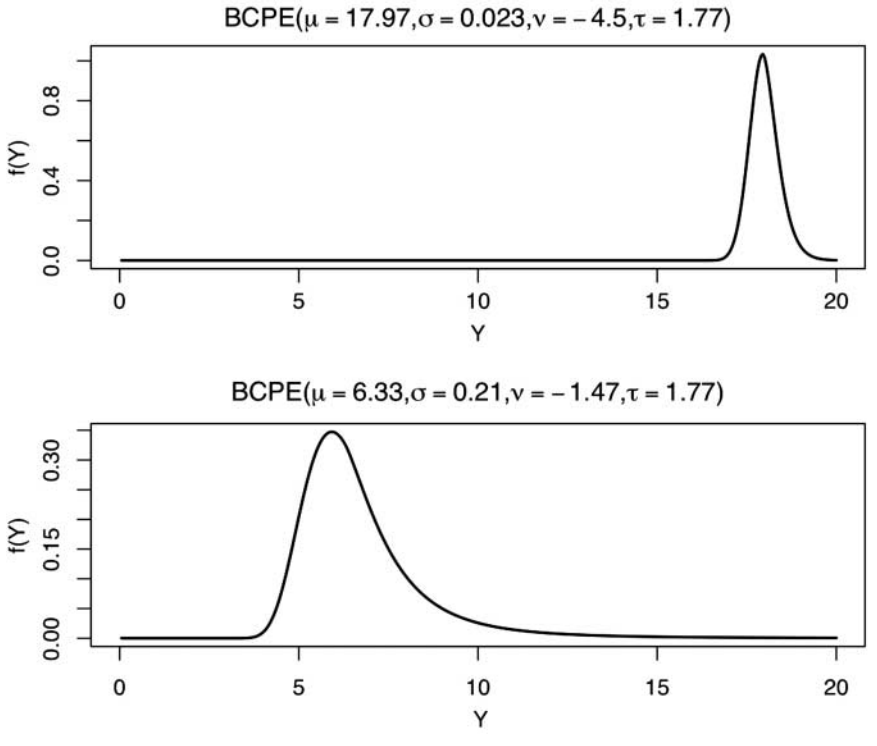


Figure 6. The pdfs for two films in the 1990s.

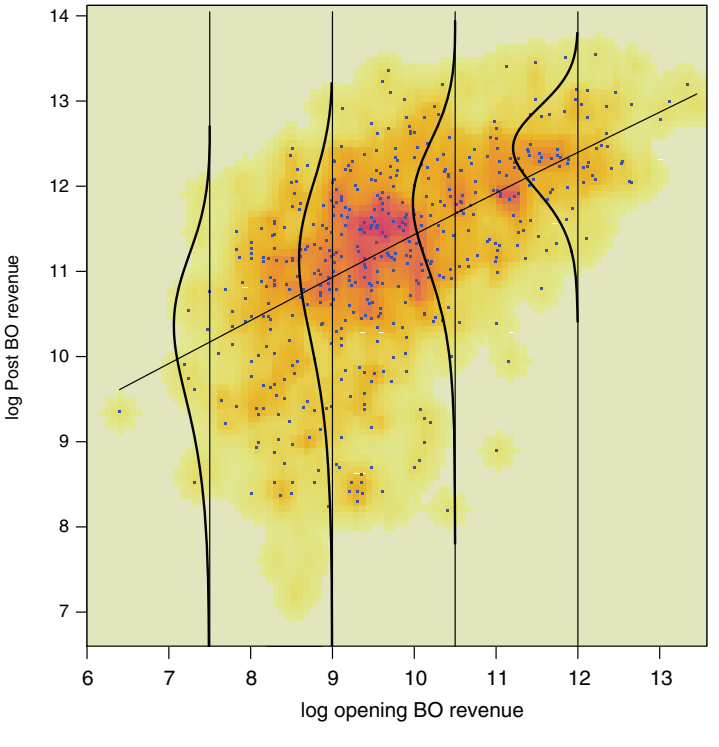


Figure 7. The fitted BCPE distribution to the 1930s' data.

## 5. Summary and conclusions

Using the flexibility of the GAMLSS framework and a wide range of model selection facilities and diagnostic procedures, we can better explore the statistical properties of the box-office revenues and revise the dominant modelling paradigm of the PLM distribution model.

The BCPE distribution model was identified by bringing a number of statistical techniques under a single umbrella, namely

- the flexibility of the GAMLSS framework for regression type of models for the distribution parameters $\mu$, $\sigma$, $\nu$ and $\tau$;
- the flexibility of the BCPE distribution to address the consequential difficulty in modelling the dispersion, skewness and kurtosis of film revenues where the BCPE distribution parameters $\mu$, $\sigma$, $\nu$ and $\tau$ may be interpreted as relating to location (median), scale (approximate coefficient of variation), skewness (transformation to symmetry) and kurtosis (power exponential parameter), respectively;
- model selection strategies for the regression type of models for the BCPE distribution parameters;
- smoothing of P-splines within GAMLSS by setting the problem as a random effect model;
- use of diagnostic plots to visualise how well a model fits the data, to find locations at which the fit can be improved and to compare different fits.

By applying the semi-parametric GAMLSS model $\mathcal{M}$ based on the BCPE distribution, we re-examined the argument of *infinite variance* and refute the claim that in the film business *nobody knows anything*. Instead, we argue that *somebody knows something*, at least in probabilistic terms.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012

## References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control 19 (1974), pp. 716–723.
[2] N.E. Breslow and D.G. Clayton, *Approximate inference in generalized linear mixed models*, J. Amer. Statist. Assoc. 88 (1993), pp. 9–25.
[3] S. van Buuren and M. Fredriks, *Worm plot: A simple diagnostic device for modelling growth reference curves*, Stat. Med. 20 (2001), pp. 1259–1277.
[4] T.J. Cole and P.J. Green, *Smoothing reference centile curves: The lms method and penalized likelihood*, Stat. Med. 11 (1992), pp. 1305–1319.
[5] A. De Vany and W.D. Walls, *Bose–Einstein dynamics and adaptive contracting in the motion picture industry*, Econ. J. 106 (1996), pp. 1493–1514.
[6] A. De Vany and W.D. Walls, *Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar*, J. Econom. Dynam. Control 28 (2004), pp. 1035–1057.
[7] P.K. Dunn and G.K. Smyth, *Randomised quantile residuals*, J. Comput. Graph. Statist. 5 (1996), pp. 236–244.
[8] P.H.C. Eilers and B.D. Marx, *Flexible smoothing with b-splines and penalties (with comments and rejoinder)*, Statist. Sci. 11 (1996), pp. 89–121.
[9] T. Haavelmo, *The statistical implications of a system of simultaneous equations*, Econometrica 11 (1943), pp. 1–12.
[10] T. Haavelmo, *The probability approach in econometrics*, Supplement to Econometrica 12 (1944).
[11] T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, London, 1990.
[12] Y. Lee, J. Nelder, and Y. Pawitan, *Generalized linear models with random effects: Unified analysis via h-likelihood*, Chapman and Hall, Boca Raton, 2006.
[13] B. Mandelbrot, *New methods in statistical economics*, J. Political Econ. 71 (1963), pp. 421–440.
[14] B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer, New York, 1997.
[15] J.A. Nelder and R.W.M. Wedderburn, *Generalized linear models*, J. R. Stat. Soc. A. 135 (1972), pp. 370–384.
[16] D.B. Nelson, *Conditional heteroskedasticity in asset returns: A new approach*, Econometrica 59 (1991), pp. 347–370.
[17] M. Pokorny and J. Sedgwick, *Profitability trends in hollywood: 1929 to 1999: Somebody must know something*, Econ. Hist. Rev. 63 (2010), pp. 56–84.
[18] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. Available at http://www.R-project.org.

[19] A.E. Raftery, *Approximate Bayes factors and accounting for model uncertainty in generalised linear models*, Biometrika 83 (1996), pp. 251–266.
[20] A.E. Raftery, *Bayes factors and bic, comment on a critique of the Bayesian information criterion for model selection*, Sociol. Methods Res. 27 (1999), pp. 411–427.
[21] R.A. Rigby and D.M. Stasinopoulos, *Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution*, Stat. Med. 23 (2004), pp. 3053–3076.
[22] R.A. Rigby and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape (with discussion)*, Appl. Statist. 54 (2005), pp. 507–554.
[23] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Vol. 104, Chapman and Hall, London, 2005.
[24] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6 (1978), pp. 461–464.
[25] W.D. Walls, *Increasing returns to information: Evidence from the Hong Kong movie market*, Appl. Econ. Lett. 4 (1997), pp. 187–190.
[26] W.D. Walls, *Modelling heavy tails and skewness in film returns*, Appl. Financ. Econ. 15 (2005), pp. 1181–1188.
[27] S. Wood, *Generalized Additive Models: An Introduction with R*, Chapman and Hall, London, 2006.

Downloaded by [Vlasios Voudouris] at 22:17 06 January 2012