Modelling electricity power cuts in the UK

Owen Dafydd Jones*

(Received 6 June 2006; revised 21 December 2006)

Abstract

We consider a compound Poisson model for electricity power cuts. Cuts occur at rate λ and we associate with the ith cut a duration L_i and size C_i, where L_i and C_i are heavy tailed and positively correlated. Development of the model is complicated by the fact that we have no direct observations of (L_i, C_i). Rather, if N is the number of power cuts in a year, we have observations of $\sum_{i=1}^{N} C_i L_i$ and $\sum_{i=1}^{N} C_i$. This necessitates the use of a parsimonious model for (L_i, C_i), and we base ours on the Pareto distribution. To fit the model we apply kernel density estimation to simulated data to obtain estimates of the likelihood, which we then maximise using stochastic optimisation.

*Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia. O.D.Jones@ms.unimelb.edu.au

See http://anziamj.austms.org.au/V47EMAC2005/Jones for this article, Published March 8, 2007. ISSN 1446-8735

Contents

1 Introduction

In what follows we describe and fit a spatio-temporal model of power cuts in the UK. A fundamental restriction on the modelling process is that the only data available to fit the model is publicly available data provided by the UK electricity supply regulator OFFER (Office of Electricity Regulation). Data reported by OFFER is aggregated over large spatial and temporal regions, yet we want to be able to use our model to simulate power cuts in detail. (OFFER divides the UK into fourteen regions and reports annually.) Accordingly we seek a parsimonious parametric model, which will necessarily make some simplifying assumptions about the process of power cuts.

From a qualitative understanding of power cuts and an understanding that, for the purposes of application, large power cuts are more important than median power cuts, it was determined that the duration and area of effect of a power cut should have heavy tailed distributions and that they

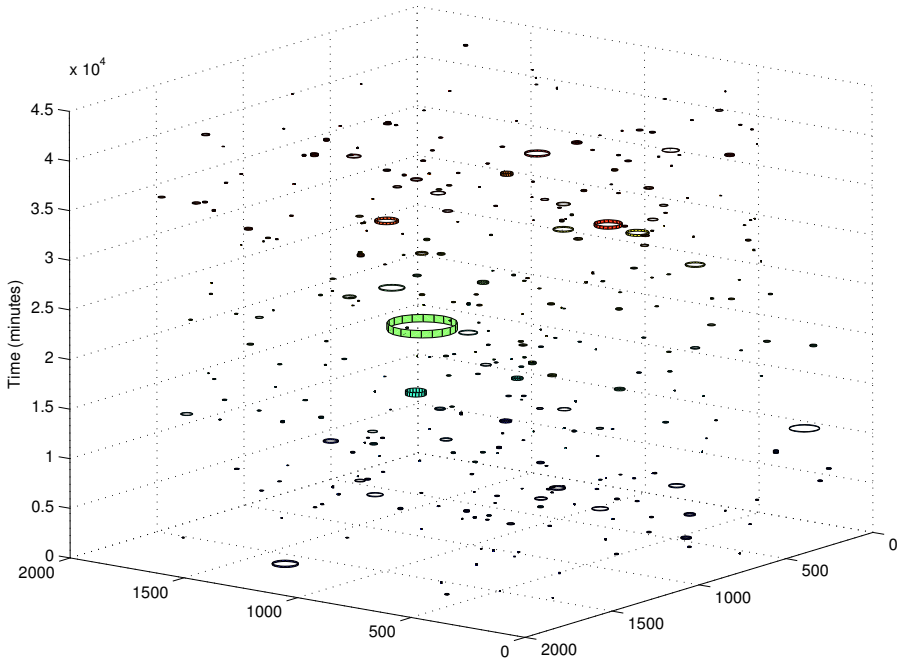

FIGURE 1: A realisation of the power cut process. Each cylinder represents a circular region affected by a power cut over a given period of time. Only cuts of duration greater than 30 minutes are shown. The spatial axes have been scaled to give one electricity customer per unit area.

should be positively correlated.

In the absence of any evidence to the contrary, it was assumed that power cuts occur over time according to a Poisson process with constant rate λ. It is likely that λ has an annual period; however, we have no way of estimating this and so ignore it. Individual power cuts are assumed to be independent and the ith cut is described by a duration L_i (in minutes) and size C_i (the number of electricity customers affected). As noted already, we have no direct observations of (L_i, C_i). Rather, if N is the number of power cuts in a year in a given geographical region and n_c is the number of electricity customers in the region, we have observations of $A := n_c^{-1} \sum_{i=1}^{N} C_i L_i$ and $B := n_c^{-1} \sum_{i=1}^{N} C_i$.

Given that a power cut has occurred, we assume that its location is randomly distributed in space with a density proportional to the population density of electricity customers over the given geographical region. That is, the time and location of power cuts forms a spatio-temporal Poisson point process, with spatial intensity proportional to the density of electricity customers and constant temporal intensity λ. The UK population density can be obtained from census data, and we assume that it is proportional to the population density of electricity customers. Given the location of a power cut, the area affected is taken to be a ball about that point with radius such that the size C_i is the number of electricity users within the ball.

Our model for (L_i, C_i) is based on the Pareto distribution. A full definition and some of its properties are given in Section 2. Figure 1 illustrates the resulting power cut process.

A non-conventional approach was required to fit our power cut model. The joint likelihood of the observed random variables A and B is not available analytically nor numerically; however, we can simulate them easily and thence use kernel density estimation to obtain an estimate of the joint likelihood, for any given set of parameter values. We then maximise the likelihood using a version of the stochastic optimisation method of Kiefer &

Wolfowitz [4]. Our approach is detailed in Section 3 together with the results of applying it to data from the region supplied by Eastern Electricity between 1990 and 1998. A discussion of our results and some related work is given in Section 4

2 The model

Our model is specified by the rate λ of power cuts and the joint distribution of the size C and duration L of a single power cut. For strictly positive parameters α, ρ_c, ρ_l, γ_c and γ_l we take

$$\begin{aligned} G &\sim \Gamma(\alpha, 1)\,, \\ X &\sim \exp(G/\rho_c) \sim \text{Pareto}(\alpha, \rho_c)\,, \\ Y &\sim \exp(G/\rho_l) \sim \text{Pareto}(\alpha, \rho_l)\,, \\ C &= X^{\gamma_c}\,, \\ L &= Y^{\gamma_l}\,, \end{aligned}$$

where X and Y are conditionally independent given G. Conditioning on G one easily obtains

$$\begin{aligned} \Pr(C > x,\, L > y) &= (1 + \rho_c^{-1} x^{1/\gamma_c} + \rho_l^{-1} y^{1/\gamma_l})^{-\alpha}\,, \\ \Pr(C > x) &= (1 + \rho_c^{-1} x^{1/\gamma_c})^{-\alpha}\,, \\ \Pr(L > y) &= (1 + \rho_l^{-1} y^{1/\gamma_l})^{-\alpha}\,. \end{aligned}$$

Thus C and L have heavy tails of order α/γ_c and α/γ_l respectively.

We obtain a further appreciation of the role played by each parameter, from the moments of C and L, when they exist. Again conditioning on G we obtain, for any $p, q > 0$ such that $\alpha > p\gamma_c + q\gamma_l$,

$$\mathbb{E}C^p L^q = \rho_c{}^{p\gamma_c} \rho_l{}^{q\gamma_l} \Gamma(p\gamma_c + 1)\Gamma(q\gamma_l + 1)\Gamma(\alpha - p\gamma_c - q\gamma_l)/\Gamma(\alpha)\,.$$

If $\alpha \leq p\gamma_c + q\gamma_l$, then the expectation is infinite. It follows that

$$\begin{aligned}
\mathbb{E}A &= \lambda {\rho_c}^{\gamma_c} {\rho_l}^{\gamma_l} \Gamma(\gamma_c+1)\Gamma(\gamma_l+1)\Gamma(\alpha-\gamma_c-\gamma_l)/(n_c\Gamma(\alpha)) \\
&\quad \text{for } \alpha > \gamma_c + \gamma_l\,, \\
\operatorname{Var} A &= \lambda {\rho_c}^{2\gamma_c} {\rho_l}^{2\gamma_l} \Gamma(2\gamma_c+1)\Gamma(2\gamma_l+1)\Gamma(\alpha-2\gamma_c-2\gamma_l)/(n_c^2\Gamma(\alpha)) \\
&\quad \text{for } \alpha > 2\gamma_c + 2\gamma_l\,, \\
\mathbb{E}B &= \lambda {\rho_c}^{\gamma_c} \Gamma(\gamma_c+1)\Gamma(\alpha-\gamma_c)/(n_c\Gamma(\alpha)) \\
&\quad \text{for } \alpha > \gamma_c\,, \\
\operatorname{Var} B &= \lambda {\rho_c}^{2\gamma_c} \Gamma(2\gamma_c+1)\Gamma(\alpha-2\gamma_c)/(n_c^2\Gamma(\alpha)) \\
&\quad \text{for } \alpha > 2\gamma_c\,, \\
\operatorname{Cov}(A,B) &= \lambda {\rho_c}^{2\gamma_c} {\rho_l}^{\gamma_l} \Gamma(2\gamma_c+1)\Gamma(\gamma_l+1)\Gamma(\alpha-2\gamma_c-\gamma_l)/(n_c^2\Gamma(\alpha)) \\
&\quad \text{for } \alpha > 2\gamma_c + \gamma_l\,.
\end{aligned}$$

For α, γ_c and γ_l outside the specified ranges these moments are infinite.

Alternatives exist to the bivariate Pareto distribution used for (X, Y), also with heavy tails and correlation, in particular the log skew t distribution and multivariate stable distribution (Nolan [9] overviewed multivariate stable distributions). An advantage of the Pareto is the particularly simple form of its joint distribution and moments and the ease with which it can be simulated. In all of these cases, the marginals for X and Y have the same tail decay rate, and so to get different decay rates for the marginals one has to introduce transformations of the form $(C, L) = (X^{\gamma_c}, Y^{\gamma_l})$, for example.

3 Model fitting

From OFFER publications [10, 11] we obtained observations of A (average supply minutes lost per customer) and B (average supply interruptions per customer) for 14 regions covering Britain, for the years 1990/91, ..., 1997/98 (8 observations per region). More recent figures are published periodically

and can be obtained from the OFFER website. We suppose that α, ρ_c, ρ_l, γ_c and γ_l remain fixed over the 14 regions, but clearly λ will vary (as does n_c, which is known).

For the purpose of demonstrating our model fitting approach we consider just the Eastern Electricity region, which covers East Anglia and parts of Greater London, with $n_c = 3,258,000$ customers. We have the following observations

Year	A	B
1990/91	76	0.76
1991/92	65	0.68
1992/93	91	0.96
1993/94	63	0.59
1994/95	94	0.65
1995/96	85	0.85
1996/97	77	0.89
1997/98	70	0.74

Sample means, variances and covariances are $\bar{a} = 77.63$, $S_A^2 = 133.70$, $\bar{b} = 0.765$, $S_B^2 = 0.016086$ and $S_{A,B} = 0.72929$.

Method of moments estimators for $\theta = (\lambda, \alpha, \rho_c, \rho_l, \gamma_c, \gamma_l)$ require a solution of the non-linear system of equations $\mathbb{E}A = \bar{a}$, $\mathbb{E}B = \bar{b}$, $\operatorname{Var} A = S_A^2$, $\operatorname{Var} B = S_B^2$ and $\operatorname{Cov}(A, B) = S_{A,B}$, subject to the constraints $\alpha > 2\gamma_c + 2\gamma_l$ and $\theta > 0$. One can solve for ρ_c and ρ_l in terms of $\mathbb{E}A$, $\mathbb{E}B$ and the other parameters, leaving three equations in four unknowns. Exact expressions for the other parameters are not readily obtainable, so we looked for a numerical solution of the corresponding constrained least-squares problem. As we have a free parameter we are able to simplify the constraint region by putting $\alpha = 2\gamma_c + 2\gamma_l + \epsilon$, where $\epsilon > 0$ is fixed. Using a local search algorithm, every attempt at finding a solution (by varying ϵ as well as the initial search position and step size) was unsuccessful and resulted in γ_c and γ_l heading to 0

without the model variances converging to the sample variances. Given the results of our maximum likelihood estimation below, a plausible explanation for our inability to match moments is that $\operatorname{Var} A$, $\operatorname{Var} B$ and $\operatorname{Cov}(A, B)$ are infinite.

The exact joint likelihood of (A, B) is in practice unobtainable. Each marginal is the mixture over possible values of N of an N-fold convolution of a Pareto-like density, where typical values of N are of the order of 7,000. However it is very easy to simulate (A, B) and given this we can use kernel density estimation (KDE) to estimate the joint likelihood. Appendix A gives sample code for simulating (A, B).

Given an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_k \in \mathbb{R}^d$, using a product form kernel, the KDE estimate of the density f at $\mathbf{x} \in \mathbb{R}^d$ is

$$\hat{f}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} \prod_{j=1}^{d} \frac{1}{\mathbf{h}(j)} K\left(\frac{\mathbf{x}(j) - \mathbf{X}_i(j)}{\mathbf{h}(j)} \right),$$

where K is the kernel, and $\mathbf{h}$ the bandwidth vector. Assume that for each j, $\mathbf{h}(j) \propto h$ for some h, then for any sensible choice of kernel (for example the normal), if f has continuous second order derivatives and square integrable third order derivatives, then we get that $\hat{f}$ has bias $\mathcal{O}(h^2)$ and variance $\mathcal{O}(k^{-1}h^{-d})$ [13]. Choosing h to minimise the Asymptotic Mean Integrated Square Error (AMISE) of $\hat{f}$ gives $h \propto k^{-1/(4+d)}$.

To estimate the log density we just use $\log \hat{f}$. Let $f_h(\mathbf{x}) = \mathbb{E}\hat{f}(\mathbf{x})$, then expanding $\log u$ about $u_0 = f_h(\mathbf{x})$ we have, at $u = \hat{f}(\mathbf{x})$,

$$\log \hat{f}(\mathbf{x}) - \log f_h(\mathbf{x}) = \frac{\hat{f}(\mathbf{x}) - f_h(\mathbf{x})}{f_h(\mathbf{x})} - \frac{(\hat{f}(\mathbf{x}) - f_h(\mathbf{x}))^2}{2 f_h(\mathbf{x})^2} + o((\hat{f}(\mathbf{x}) - f_h(\mathbf{x}))^2).$$

Thus

$$\mathbb{E} \log \hat{f}(\mathbf{x}) = \log f_h(\mathbf{x}) + \mathcal{O}(k^{-1}h^{-d}) = \log f(\mathbf{x}) + \mathcal{O}(h^2) + \mathcal{O}(k^{-1}h^{-d}).$$

Expanding $(\log u - \log f_h(\mathbf{x}))^2$ about $u_0 = f_h(\mathbf{x})$ we have, at $u = \hat{f}(\mathbf{x})$,

$$\Big(\log \hat{f}(\mathbf{x}) - \log f_h(\mathbf{x})\Big)^2 = (\hat{f}(\mathbf{x}) - f_h(\mathbf{x}))^2 + o((\hat{f}(\mathbf{x}) - f_h(\mathbf{x}))^2)\,.$$

Thus

$$\begin{aligned}
& \operatorname{Var} \log \hat{f}(\mathbf{x}) \\
\le\ & \mathbb{E}(\log \hat{f}(\mathbf{x}) - \log f(\mathbf{x}))^2 \\
=\ & \mathbb{E}\left(\log \frac{\hat{f}(\mathbf{x})}{f_h(\mathbf{x})}\right)^2 + 2\left(\log \frac{f_h(\mathbf{x})}{f(\mathbf{x})}\right)\mathbb{E}\left(\log \frac{\hat{f}(\mathbf{x})}{f_h(\mathbf{x})}\right) + \left(\log \frac{f_h(\mathbf{x})}{f(\mathbf{x})}\right)^2 \\
=\ & \mathcal{O}(k^{-1}h^{-d}) + \mathcal{O}(k^{-1}h^{2-d}) + \mathcal{O}(h^4) \\
=\ & \mathcal{O}(k^{-1}h^{-d}) + \mathcal{O}(h^4)\,.
\end{aligned}$$

In applying KDE we used code based on that of Beardah [1]. We used the standard normal kernel.

3.1 Stochastic optimisation

Since A and B inherit skewed heavy tails from C and L, rather than apply KDE to the joint density $f_{A,B}$ of A and B, we apply it to the joint density $f_{\log A,\log B}$ of $\log A$ and $\log B$. The relationship between the two densities is

$$\log f_{A,B}(x,y) = \log f_{\log A,\log B}(\log x, \log y) - \log x - \log y\,.$$

The joint density $f_{\log A,\log B}$ is smooth enough for our KDE estimates to converge as above.

Let (a_i, b_i), $i = 1,\dots,m$, be our observations of (A,B) then for $\theta = (\lambda, \gamma_c, \rho_c, \gamma_l, \rho_l, \alpha) > 0$ our approximation of the log likelihood is

$$\hat{l}(\theta) = \sum_{i=1}^{m} \Big(\log \hat{f}_{\log A,\log B}(\log a_i, \log b_i) - \log a_i - \log b_i\Big)\,,$$

where $\hat{f}_{\log A, \log B}$ is our KDE estimate, so that $\hat{l}$ has bias $\mathcal{O}(k^{-1}h^{-d}) + \mathcal{O}(h^2)$ and variance $\mathcal{O}(k^{-1}h^{-d}) + \mathcal{O}(h^4)$, as the simulation sample size $k \to \infty$ and kernel bandwidth $h \to 0$.

We maximise the log likelihood over θ using the stochastic optimisation technique of Kiefer & Wolfowitz [4, 2]. Let θ_n be the nth candidate solution then for non-negative sequences $\{a_n\}$ and $\{c_n\}$ we have for $j = 1, \ldots, d_\theta = 6$

$$\begin{aligned} \theta_{n+1}(j) &= \theta_n(j) + a_n \frac{\hat{l}(\theta_n + c_n \mathbf{e}_j) - \hat{l}(\theta_n - c_n \mathbf{e}_j)}{2c_n} \\ &= \theta_n(j) + a_n (l_j(\theta_n) + \beta_n(j) + \xi_n(j)), \end{aligned}$$

where $l_j = \partial l / \partial \theta(j)$, $\mathbf{e}_j$ is the unit vector with jth coordinate equal to 1, and the bias and error terms are

$$\begin{aligned} \beta_n(j) &= \frac{l(\theta_n + c_n \mathbf{e}_j) - l(\theta_n - c_n \mathbf{e}_j)}{2c_n} - l_j(\theta_n) \\ &\quad + \mathbb{E}\left(\frac{\hat{l}(\theta_n + c_n \mathbf{e}_j) - \hat{l}(\theta_n - c_n \mathbf{e}_j)}{2c_n} \right) - \frac{l(\theta_n + c_n \mathbf{e}_j) - l(\theta_n - c_n \mathbf{e}_j)}{2c_n}, \\ \xi_n(j) &= \frac{\hat{l}(\theta_n + c_n \mathbf{e}_j) - \hat{l}(\theta_n - c_n \mathbf{e}_j)}{2c_n} - \mathbb{E}\left(\frac{\hat{l}(\theta_n + c_n \mathbf{e}_j) - \hat{l}(\theta_n - c_n \mathbf{e}_j)}{2c_n} \right). \end{aligned}$$

In practice a_n and c_n can be allowed to depend on j.

Since l is twice continuously differentiable, we have that θ_n converges almost surely to a point of local maximum of l if the following conditions hold [6]:

1. $\sum_n a_n = \infty$ and $\sum_n a_n^2 < \infty$;
2. $\sum a_n^2 c_n^{-2} < \infty$;
3. For each j, $\mathbb{E}|\xi_n(j)|$, $c_n^2 \operatorname{Var} \xi_n(j)$ and $\beta_n(j)^2$ are bounded in n.

(More general conditions for convergence are given by Kushner & Yin [6], but are not needed here.) Let h_n be the bandwidth and k_n the sample size used to obtain the KDE estimates $\hat{l}(\theta_n \pm c_n \mathbf{e}_j)$. Taking $a_n = n^{-1}$ and $c_n = n^{-\delta}$ for $0 < \delta < 0.5$ satisfies Conditions 1 and 2. Choosing h_n and k_n to satisfy Condition 3 and minimise k_n we get, since $\mathbb{E}|\xi_n(j)| \leq 1 + \operatorname{Var} \xi_n(j)$,

$$\begin{aligned} \beta_n(j)^2 &= \mathcal{O}(k_n^{-2} h_n^{-2d} c_n^{-2}) + \mathcal{O}(k_n^{-1} h_n^{2-d} c_n^{-2}) + \mathcal{O}(h_n^4 c_n^{-2}) = \mathcal{O}(1), \\ \operatorname{Var} \xi_n(j) &= \mathcal{O}(k_n^{-1} h_n^{-d} c_n^{-2}) + \mathcal{O}(h_n^4 c_n^{-2}) = \mathcal{O}(1), \end{aligned}$$

whence

$$k_n \propto n^{(2+d/2)\delta} \quad \text{and} \quad h_n \propto n^{-\delta/2} \propto k_n^{-1/(4+d)}.$$

Note that this relation between h_n and k_n is the same as that given by minimising the AMISE of $\hat{f}$.

Note that in practice the asymptotic requirements for a_n, c_n, h_n and k_n, which are needed to ensure that the algorithm eventually converges, are much less important than their initial values and in practice they can usually be kept constant.

The speed of convergence of the stochastic optimisation algorithm was significantly improved by using the "common random numbers" variance reduction technique [5]. That is, at each iteration the same sequence of pseudo random numbers was used to generate $\hat{l}(\theta_n \pm c_n \mathbf{e}_j)$ for $j = 1, \ldots, d_\theta$, which has the effect of reducing the variance of the $\xi_n(j)$. Appendix A indicates how to modify our simulation code to take advantage of common random numbers.

3.2 Results

Having a small data set meant that the log likelihood l was very close to $-\infty$ for most values of θ and quite flat for the remainder, so it was necessary to take small step sizes. Also l is much more sensitive to changes in α, γ_c and γ_l than to changes in ρ_c, ρ_l and λ, which necessitated a coordinate scaling of θ.

An initial value θ_0 was obtained by putting $\lambda = 10,000$ (based on some auxiliary information of poor quality), $\rho_c = \rho_l = 1$, $\alpha = 3$ and then solving $\mathbb{E}A = \bar{a}$ and $\mathbb{E}B = \bar{b}$ to find γ_c and γ_l. From this point the stochastic optimisation algorithm converged slowly but steadily. Our final values for the components of θ were

$$\begin{array}{rclrclrcl} \hat{\lambda} &=& 7,122\,, & \hat{\gamma_c} &=& 1.9955\,, & \hat{\rho_c} &=& 16.71\,, \\ \hat{\gamma_l} &=& 0.6255\,, & \hat{\rho_l} &=& 838.60\,, & \hat{\alpha} &=& 2.7730\,. \end{array}$$

This value of λ corresponds to approximately 20 power cuts per day across the Eastern Electricity region (covering all of East Anglia and parts of London). This is reasonable, bearing in mind that most of these are of short duration. Under this model $\mathbb{E}A$ and $\mathbb{E}B$ are finite but $\operatorname{Var} A$, $\operatorname{Var} B$ and $\operatorname{Cov}(A, B)$ are not. We have $\mathbb{E}A = 271.50$ and $\mathbb{E}B = 0.872$. These are the correct order of magnitude, although $\mathbb{E}A$ is nearly four times greater than the observed $\bar{a}$. The explanation is in the heavy tails of A: the observed sample does not include any extreme values and thus $\bar{a}$ gives an underestimate of $\mathbb{E}A$.

Having obtained $\hat{\theta}$, we use $\log \hat{f}$ to estimate the Fisher information matrix. We approximated the second order partial derivatives of $\log \hat{f}$ using differences, where common random numbers were used to estimate $\log \hat{f}$ for different values of θ. Note that when using KDE to estimate derivatives rather than the function itself, you should use a wider bandwidth. The estimated standard deviations for our estimators were

$$\begin{array}{rclrclrcl} \hat{\sigma}_{\lambda} &=& 131.9\,, & \hat{\sigma}_{\gamma_c} &=& 0.0931\,, & \hat{\sigma}_{\rho_c} &=& 21.00\,, \\ \hat{\sigma}_{\gamma_l} &=& 0.0655\,, & \hat{\sigma}_{\rho_l} &=& 10,789\,, & \hat{\sigma}_{\alpha} &=& 0.0481\,. \end{array}$$

Our estimates for the rate λ of power cuts and the tail decay rates, given by α, γ_c and γ_l, are not too bad; but our estimates for ρ_c and ρ_l, which determine how the distributions of C and L are shifted, are unreliable.

An alternative method for estimating the information matrix is to bootstrap (studentised) or jackknife from the original data set. This approach is relatively time consuming, as each run of the stochastic optimisation algorithm takes some time.

Figure 2 plots (approximate) contours of $f_{\log A, \log B}$ with the observed values of $(\log A, \log B)$ superimposed. The fit is reasonable although there is clearly room for improvement. The problem is that we cannot increase the variation of B without also increasing the variation of A. Considering again the physical rationale for our model, we have good cause to believe there is positive correlation between C (number of customers affected) and L (duration) for large power cuts, because a determining factor is the quantity of resources available for repairing the problem. This argument does not apply to small power cuts so we have no firm basis for supposing C and L are positively correlated in this case. Clearly the correlation structure of our model is not flexible enough to model both of these regimes at once. Without direct observations of C and L it is difficult to draw further conclusions about the disparities between model and observations.

4 Discussion

The method used to fit our compound Poisson model for power cuts has the advantage of being very easy to implement, although it is not particularly fast and is subject to some uncertainty. It is also very flexible, in the sense that it is very easy to incorporate additional information. For example, [10] includes a single observation of the random variable $Z = N^{-1} \sum_{i=1}^{N} I\{L_i \geq 180\}$ for the year 1997/98. Given that we already simulate $L_1, \ldots, L_N$ in order to simulate A, very little extra effort is required to of simulate Z, from which we can easily estimate the joint density of (A, B, Z) and thereby include our observation of Z in the estimated likelihood.

A potential limitation of our method stems from the well known problem that kernel density estimation performs poorly in higher dimensions. In practice KDE is rarely used for the joint density of more than five or six variables, as the sample size required to get a reasonable estimate becomes too large as the dimension increases.

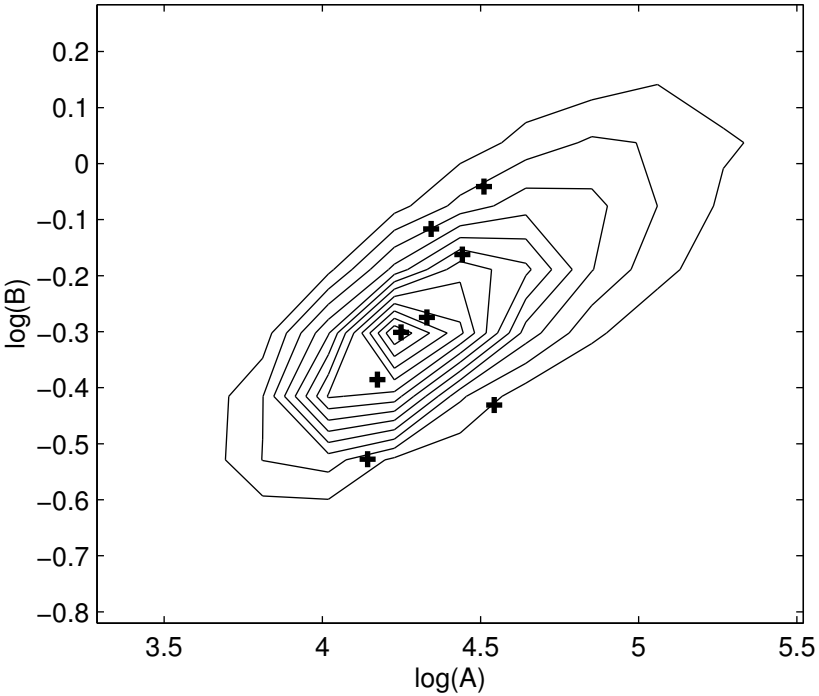

FIGURE 2: (Estimated) contour plot of the joint density of $(\log A, \log B)$ for the maximum likelihood estimate of θ. The observed sample points have been overlaid for comparison.

Our use of simulation to estimate the log likelihood is similar in some respects to the "method of simulated moments" seen in economics research [8, 12]. There as here, it is supposed that one has a parameterised model that is relatively easy to simulate but hard to analyse otherwise. As the name suggests, rather than using the likelihood, the method of simulated moments seeks to match the sample moments of the simulated process with observed sample moments, using numerical optimisation techniques to find the optimal parameter values. An important difference between the two approaches is that the method of simulated moments uses only a single random sample, rather than generating a new sample each time we try new parameter values. That is, we suppose that the simulation approximation, $\hat{g}(\theta)$ say, can been written as $G(\theta, \mathbf{X})$, where the simulation sample $\mathbf{X}$ does not depend on the parameters θ, so that you can optimise over θ without repeatedly simulating $\mathbf{X}$. In our case it is possible to formulate $\hat{l}(\theta)$ in this way, by thinking of it as a function of a stream of pseudo random numbers $\mathbf{U}$ (simulated i.i.d. uniform random variables). By resetting the seed used to generate $\mathbf{U}$ to the same value each time $\hat{l}(\theta)$ is generated, we can view it as a deterministic function of θ and thus use a deterministic local search method to maximise it, rather than stochastic optimisation. The disadvantage of this approach is that the solution is now a (biased) random variable, with variation dependent on the simulation sample size.

Lee [7] also considered simulated maximum likelihood, though not using kernel methods nor stochastic optimisation. He assumes that he has an unbiased estimator $\hat{f}$ of f then considers the bias and variance of $\log \hat{f}$.

Our method for estimating $\log f$ is similar to kernel estimators of the entropy [3]. In both cases we have a sum $\sum_{i=1}^{m} \log \hat{f}(\mathbf{x}_i)$ where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are our observations. The difference is that our $\hat{f}$ is formed from an independent simulated sample, whereas the entropy estimator $\hat{f}$ is a kernel density estimate of f formed using the same sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Not surprisingly, this dependency between $\hat{f}$ and the sample points at which it is evaluated makes analysis of the entropy estimator relatively difficult.

Finally we mention a rather different route to estimating the likelihood, using characteristic functions. Conditioning on N, one can easily show that

$$\mathbb{E}e^{iuA+ivB} = e^{-\lambda(1-\phi((u+v)/n_c))},$$

where ϕ is the characteristic function of $C(1+L)$. ϕ can be calculated numerically, whence the joint density of (A, B) is obtained by applying the inverse fast Fourier transform to $\mathbb{E}e^{iuA+ivB}$. While more technical, this approach should prove less numerically intensive than the stochastic optimisation method. Unfortunately it is quite specific to the form of A and B, and cannot be extended to include Z, for example.

A Simulation code

The following code was used to simulate (A, B), using Matlab Release 12.1 and the Statistics Toolbox [14]. `lambda`, `alpha`, `rhoC`, `gammaC`, `rhoL` and `gammaL` are the parameters, `n` is the number of electricity customers and `N`, `G`, `C`, `L`, `A` and `B` correspond to the variables with the same names defined in Sections 1 and 2, noting that `G`, `C` and `L` are vector valued.

```
N = poissrnd(lambda);
G = gamrnd(alpha,1,[1,N]);
C = exprnd(rhoC./G).^gammaC;
L = exprnd(rhoL./G).^gammaL;
A = sum(C.*L)/n;
B = sum(C)/n;
```

Since `gamrnd` uses a rejection method for simulating gamma random variables and thus uses a non-constant number of pseudo random numbers, when applying the common random variables technique we reformed our algorithm so that the same pseudo random numbers would be used to generate `C` and `L` each time:

```
N = poissrnd(lambda);
U1 = rand(1,N);
U2 = rand(1,N);
G = gamrnd(alpha,1,[1,N]);
C = (-log(U1).*rhoC./G).^gammaC;
L = (-log(U2).*rhoL./G).^gammaL;
A = sum(C.*L)/n;
B = sum(C)/n;
```

References

[1] Beardah, C. C., Kernel Density Estimation "Toolbox", Version 1.1. Dept. of Maths, Stats and OR, The Nottingham Trent University, http://euler.ntu.ac.uk/maths.html. 22nd October 1996. C611

[2] Blum, J. R., Multidimensional stochastic approximation methods. *Ann. Math. Statist.* 25, pp. 737–744, 1954. C612

[3] Hall, P. and Morton, S. C., On the estimation of entropy. *Ann. Inst. Statist. Math.* 45, pp. 69–88, 1993. doi:10.1007/BF00773669 C617

[4] Kiefer, J. and Wolfowitz, J., Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23, pp. 462–466, 1952. C607, C612

[5] Kleinman, N. L., Spall, J. C. and Naiman, D. Q., Simulation-based optimization with stochastic approximation using common random numbers. *Management Science* 45, pp. 1570–1578, 1999. C613

[6] Kushner, H. J. and Yin, G. G., *Stochastic Approximation Algorithms and Applications*, Springer–Verlag, 1997. C612, C613

[7] Lee, L-F., Statistical inference with simulated likelihood functions. *Economic Theory* 15, pp. 337–360, 1999. doi:10.1017/S0266466699153039 C617

[8] McFadden, D., A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57, pp. 995–1026, 1989. C617

[9] Nolan, J. P., Multivariate stable distributions: approximation, estimation, simulation and identification. In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Adler, R. J., Feldman, R. E. and Taqqu, M. S. eds., pp. 509–525. Birkhäuser, 1998. C608

[10] OffER. Report on Distribution and Transmission System Performance 1997/98. Office of Electricity Regulation, Hagley House, Hagley Road, Birmingham B16 8QG, U.K. http://www.ofgem.gov.uk/ofgem/whats-new/archive.jsp. November 1998. C608, C615

[11] OffER. Review of Public Electricity Suppliers 1998–2000: Distribution Price Control Review Consultation Paper. Office of Electricity Regulation, Hagley House, Hagley Road, Birmingham B16 8QG, U.K. http://www.ofgem.gov.uk/ofgem/whats-new/archive.jsp. May 1999. C608

[12] Pakes, A. and Pollard, P., Simulation and the asymptotics of optimization estimators. *Econometrica* 57, pp. 1027–1057, 1989. C617

[13] Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, 1992. C610

[14] The MathWorks, Inc., Matlab Version 6.1.0.450 Release 12.1. 18 May 2001. C618