Electricity time series: stylized facts and model estimation

Peter Tankov[1]

[1]Université Paris-Diderot

Séminaire du Laboratoire Finance des Marchés d'Energies
Paris, May 14, 2007

Outline of the talk

- Description of data
- Stylized features of electricity prices
- Literature review
- Spike detection
- Model estimation

Data sets: daily or hourly

"Spot" hourly prices are day ahead prices determined simultaneously for all 24 hours of the next day

$\Rightarrow$ No causality relationship between hourly prices of the same day

$\Rightarrow$ No a priori reasons to model spot price as an hourly series

It is preferable to model the price X_t^h for day t and hour h as

$$X_t^h = Y_t f(t,h) + \varepsilon_t^h,$$

- Y_t is the common factor (average daily price)
- $f(t,h)$ is a slowly varying daily pattern
- ε_t^h is a white noise.

All interesting statistical features are present in average daily price

PCA shows that even a constant daily pattern $f(t,h) = f(h)$ explains 70% of variance.

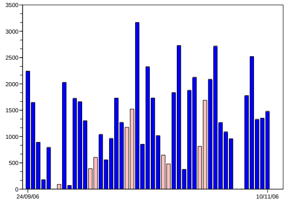
$\Rightarrow$In this talk we concentrate on average daily prices

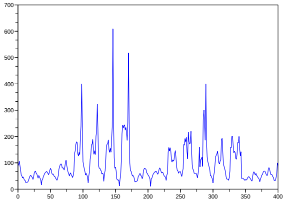
The weekend effect

On the weekends, prices and trading volume are low

They introduce a lot of seasonality but no interesting statistical features (e.g., no spikes during weekends)

⇒ to make deseasoning easier and concentrate on statistical aspects, weekends are removed.





Description of data sets

In this work, we study average daily prices excluding weekends for 7 series:

- Dow-Jones California-Oregon border index (COB)
- Dow-Jones Mead/Marketplace index (MEAD)
- European Energy Exchange (EEX)
- Amsterdam Power Exchange (APX)
- United Kingdom Power Exchange (UKPX)
- Nord Pool system price (NP)
- Powernext exchange (PN)

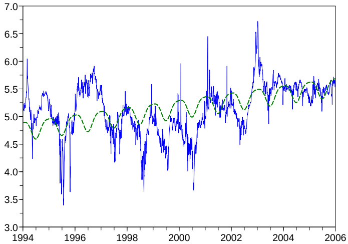
Seasonality

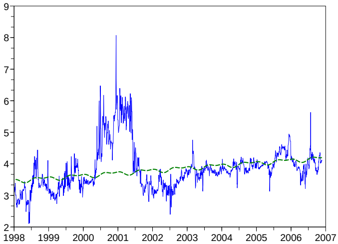
All 6 series exhibit a (strong mostly positive) linear trend and a (weak) seasonality.

$$f(t) = a + bt + c_1 \sin(2\pi t) + c_2 \cos(2\pi t) + d_1 \sin(4\pi t) + d_2 \cos(4\pi t)$$

Series	Mean growth rate	Affine trend R^2	Seasonal trend R^2
COB	8%	0.0990	0.1031
MEAD	−5%	0.0274	0.0385
EEX	17%	0.5160	0.5373
APX	9%	0.1378	0.1715
UKPX	20%	0.5615	0.5900
NP	7%	0.2504	0.3106
PN	15%	0.2966	0.3327

Seasonality





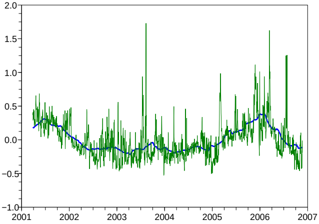
Left: Nord Pool series (strongest seasonal effect). Right: COB series (weakest seasonality).

Stationarity and mean reversion

- For all 7 series the non-stationarity hypothesis is rejected by the DF test at 1% level (test statistics greater than 29.5).

Series	APX	COB	EEX	MEAD	NP	PN	UKPX
DF test statistics	466	59.4	483	43.2	66.4	294	284

- The prices revert to a slowly varying stochastic mean level.



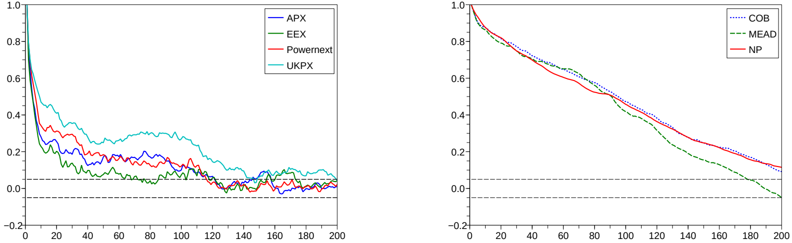
UKPX series

Autocorrelation structure

Two very different autocorrelation patterns:

- UKPX, EEX, APX and Powernext: fast decay to 20-40% followed by slow decay.
- COB, MEAD and Nordpool: very slow decay, almost non-stationary series.

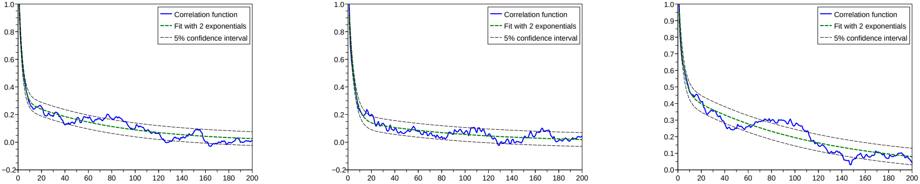
→ differences in market organization and generation facilities



Multiscale autocorrelation

- For APX, EEX, Powernext and UKPX series, the autocorrelation structure is described precisely by

$$\rho(h) = w_1 e^{-h/\lambda_1} + w_2 e^{-h/\lambda_2}.$$



From left to right: APX, EEX, UKPX

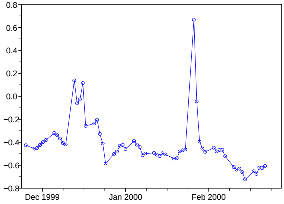
Multiscale autocorrelation

	APX	EEX	Powernext	UKPX
λ_1	2.9	4.3	2.9	3.7
λ_2	81.5	94.3	62.1	112.6

This correlation structure arises in a model where the price is a sum of two independent mean-reverting components, with fast and slow mean reversion.

Spikes

- Spikes: fast upward movements followed by quick return to initial level.
- Fundamental feature of electricity prices, due to non-storable nature of this commodity.



Non-gaussian return distribution

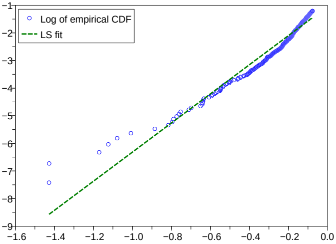
- Spikes cannot arise in the Gaussian framework.
- The return distributions are strongly leptokurtic and positively skewed (spikes are mostly positive).

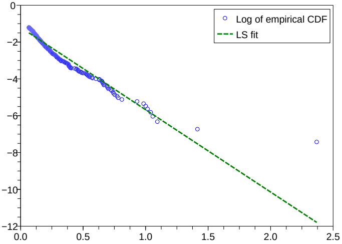
Series	APX	COB	EEX	MEAD	NP	UKPX
Skewness	0.11	0.163	0.641	0.025	0.431	0.801
Excess kurtosis	14.8	15.4	12.7	9.5	29.0	15.0

- Excess kurtosis for S&P 500 ~ 3.

Non-gaussian return distribution

- Tails of the return distribution may be fatter than exponential





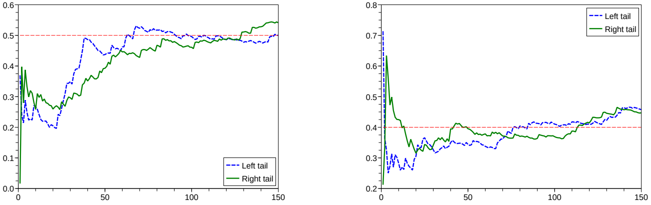
Left graph: $\log F_n(x)$; right graph: $\log(1 - F_n(x))$ where F_n is the empirical CDF of EEX returns.

Non-gaussian return distribution

Testing power law behavior $1 - F(x) \sim x^{-\alpha} L(x)$: the Hill plot

$$H_{k,n} := \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}}$$

can be used to estimate the tail index $\gamma = \alpha^{-1}$



Hill plots for MEAD (left) and UKPX (right). For all 7 series, the Hill plot stabilizes for $0.3 < \gamma < 0.5$.

Structural models

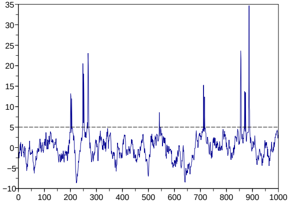
Kanamura, Ohashi (2004)

- Price is obtained by matching stochastic demand with deterministic supply curve.

$D_t = \overline{D_t} + X_t$ (seasonal effect)

$dX_t = (\mu - \lambda X_t)dt + \sigma dW_t.$ (stochastic part)

$P_t = (a_1 + b_1 D_t)1_{D_t \leq D_0} + (a_1 + b_1 D_t)1_{D_t > D_0}$ (hockey stick profile)



- No stochastic base level
- No multiscale autocorrelation
- Difficult to calibrate

Markov models

Geman and Roncoroni (2006)

$$dP_t = \theta(\mu_t - P_t) + \sigma dW_t + h(t)dJ_t$$

The jump direction and intensity are level-dependent, "jump-reversion"

- No stochastic base level

Regime-switching models

Deng (1999), Weron (2005)

- An unobservable 2-state Markov chain determines the transition from "base regime" to "spike regime" with greater volatility and faster mean reversion.

$$dP_t = \theta^1(\mu_t - P_t) + \sigma^1 dW_t \quad \text{(base regime)}$$
$$dP_t = \theta^2(\mu_t - P_t) + \sigma^2 dW_t \quad \text{(spike regime)}$$

- Nonlinear dynamics makes estimation and pricing difficult
- No stochastic base level (in the spike regime, the process quickly reverts to seasonal mean and not to base level).

Multifactor models

- Log-price is a sum of independent Ornstein-Uhlenbeck components: Villaplana (2003), Deng, Jiang (2005), Benth, Kallsen, Meyer-Brandis (2006)

$$X(t) = \sum_{i=1}^{n} w_i Y_i(t)$$

where

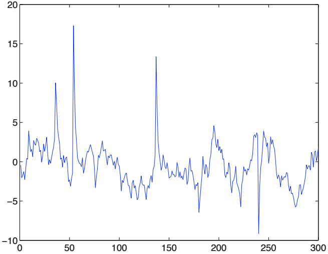
$$dY_i(t) = -\lambda_i^{-1} Y_i(t) dt + \sigma_i dL_i(t) \quad Y_i(0) = y_i$$

and processes $L_i(t)$ are independent Lévy processes.
In this model, the autocorrelation function is

$$\rho(h) = \frac{\sum_{i=1}^{n} w_i^2 e^{-h/\lambda_i}}{\sum_{i=1}^{n} w_i^2}$$

Multifactor models

- In practice, one can take $n = 2$, L_1 a Brownian motion and L_2 a compound Poisson process.



Discrete version of the model

$$X(t) = Y_1(t) + Y_2(t)$$
$$Y_1(t) = e^{-1/\lambda_1} Y_1(t-1) + \varepsilon_1(t)$$
$$Y_2(t) = e^{-1/\lambda_2} Y_2(t-1) + \varepsilon_2(t)$$

Spike detection

Estimation becomes easy if the two components are separated

- Threshold methods: detecting jumps, not spikes
- Nonlinear filtering methods
- Methods from non-parametric statistics

Nonlinear filtering

Filtering problem: estimate the spike component via

$$E[f(Y_2(t))|X_1, \dots, X_t]$$

In the non-Gaussian framework explicit Kallman filter cannot be used and must use Monte Carlo methods (particle filters). However

- Filters are easy to design when parameters are known;
- Rare events such as spikes lead to sample impoverishment;
- Sequential filtering makes less sense when complete series is available.

Methods from nonparametric statistics

Idea: treat the spike part $Y_2(t)$ as deterministic data and the base part $Y_2(t)$ as random (autoregressive) noise.

$$X(t) = Y_1(t) + f(t).$$

$$f(t) = \sum_{i=1}^{M} \alpha_i 1_{t \geq \tau_i} e^{-(t-\tau_i)/\lambda_2}.$$

The ML estimator of (α_i, τ_i) is

$$(\alpha_i, \tau_i) = \arg\inf \sum_{t=1}^{N} (\Delta X(t) - \Delta f(t))^2$$

$$\Delta X(t) = X(t) - e^{-1/\lambda_1} X(t-1), \qquad \Delta f(t) = f(t) - e^{-1/\lambda_1} f(t-1).$$

$\Rightarrow$ complexity N^M.

How to place one spike?

If λ_1 and λ_2 are known (from autocorrelation function)

$$f(t) = 1_{t \geq \tau} e^{-(t-\tau)/\lambda_2}.$$

The ML estimator of α and τ is

$$(\alpha^*, \tau^*) = \arg\inf \sum (\Delta X(t) - \alpha \Delta f(t))^2$$

$$\alpha^* = \frac{\sum \Delta X(t) \Delta f(t)}{\Delta f(t)^2}, \qquad \tau^* = \arg\sup \frac{(\sum \Delta X(t) \Delta f(t))^2}{\Delta f(t)^2},$$

Hard thresholding

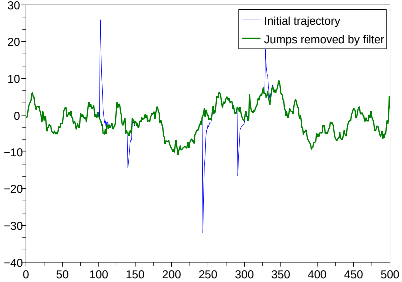
- Place M spikes one by one

$$X^{(0)} = X, \qquad X^{(n+1)} = X^{(n)} - f^{(n)}, \qquad f^{(n)}(t) = \alpha^n 1_{t \geq \tau^n} e^{-(t-\tau^n)/\lambda_2}.$$

$$\alpha^n = \frac{\sum \Delta X^{(n)}(t) \Delta f(t)}{\Delta f(t)^2}, \qquad \tau^n = \arg\sup \frac{\left(\sum \Delta X^{(n)}(t) \Delta f(t)\right)^2}{\Delta f(t)^2},$$

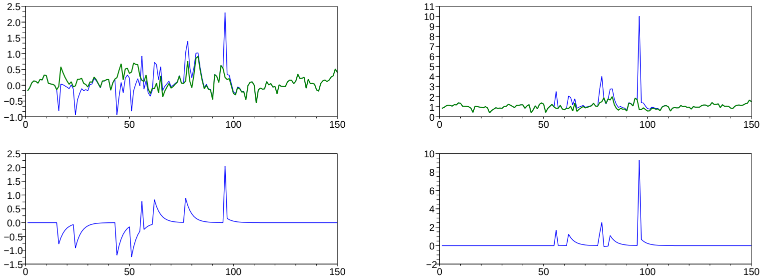
- Complexity MN
- The procedure stops when spike size becomes small
- The method works well for rare non-intersecting spikes

Hard thresholding



- Performance on simulated data.

Hard thresholding



- Performance on Powernext data

Left: log-price: both positive and negative spikes are present.
Right: price: mostly positive spikes are present.

Hard thresholding

- The hard thresholding algorithm requires λ_1 (base decay time) and λ_2 (spike decay time) as inputs. From autocorrelation function, $\lambda_1 \sim 50 - 100$ and $\lambda_2 \sim 2.5 - 4$.
- The results are not sensitive to λ_1 in the range $10 - 100$.
- Taking $\lambda_2 \sim 1 - 2$ leads to better performance:
  - Spikes decay faster than predicted by the autocorrelation function.
  - The hard thresholding algorithm works better for non-interacting spikes.

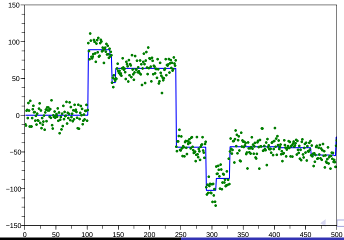
To improve performance for interacting spikes:

- Global optimization using genetic algorithms.
- Dynamic programming approach: the Potts filter.

The Potts filter

The Potts filter (Winkler and Liebscher '02): estimation of a piecewise-constant signal x from noisy data y by penalized least squares:

$$x = \arg\inf \left(\gamma |\{t : x_{t-1} \neq x_t\}| + \sum_t (y_t - x_t)^2 \right)$$



The Potts filter

Solution in $O(N^2)$ via dynamic programming:

$$B(n) = \min_{1\le r\le n-1} \left(B(r) + \min_{\mu\in\mathbb{R}} H_{[r+1,n]}(\mu) \right),$$

where

$$H_{[a,b]}(\mu) = \gamma + \sum_{a\le t\le b} (y_t - \mu)^2.$$

is the cost of adding an interval to the partition and

$$B(n) = \min \left(\gamma |\{t \le n : x_{t-1} \ne x_t\}| + \sum_{t\le n} (y_t - x_t)^2 \right)$$

is the solution on $[1, n]$.

Modified Potts filter for spike detection

Idea: replace the cost of adding an interval with the cost of adding a spike:

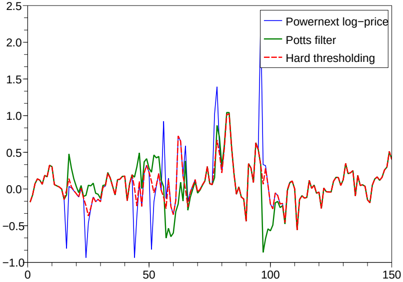
$$\min_{\mu} H_{[a,b]}(\mu) \mapsto \gamma + \inf_{\alpha} \sum_{t=a+1}^{b} (\Delta y_t - \alpha \Delta e^{-(t-a)/\lambda_2})^2$$

with

$$\Delta z_t = z_t - e^{-1/\lambda_1} z_{t-1}.$$

- The filter is designed for detecting discontinuities so it detects jumps as well as spikes.

Modified Potts filter for spike detection

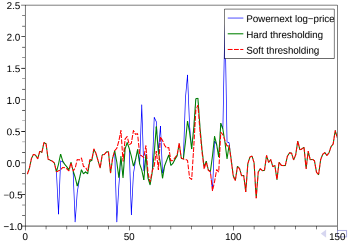


Soft thresholding

- Penalization by the L^1 norm of spike sizes

$$\inf \sum_{t=1}^{N} (\Delta X(t) - \Delta f(t))^2 + \gamma \sum_{i=1}^{n} |\alpha_i|$$

- Can be approximated with a linear programming program
- Better treatment of adjacent spikes $\Rightarrow$ more spikes can be identified



Model estimation

Once the two components have been separated, model estimation can be performed separately on each component.

- The base signal is described by an AR(1) model with gaussian or NIG returns, estimation by maximum likelihood.
- The spike sizes are fitted by a one-parameter family (Pareto or exponential).
- The spike intensity is $\lambda = \frac{\text{number of spike detected}}{\text{time period}}$.
- The threshold is fixed by comparing the spike amplitude with the overall noise level.

Case study: EEX series

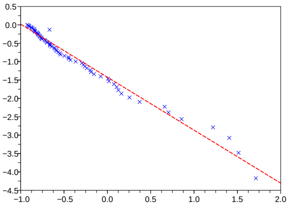
- Jumps are removed by hard thresholding from deseasonalized price series.
- The threshold level is set to one standard deviation of price increments (0.38) $\longrightarrow$ 65 spikes are detected in 6.5 years of data.
- After removing spikes, the increments of the base series have standard deviation of 0.15, skewness of 0.07 (down from 2.6) and kurtosis of 0.65 (down from 113).
  $\Rightarrow$ base signal can be described by AR(1) with Gaussian increments.
  Estimation gives $\rho = 0.86$ and $\sigma = 0.142$.

Case study: EEX series

The spike size distribution is well described by the Pareto law

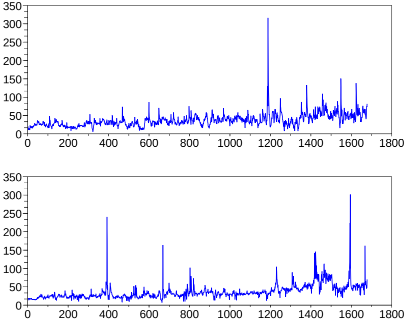
$$P(X > x) = \left(\frac{x}{x_0}\right)^{-\alpha}$$

with $x_0 = 0.37$ and $\alpha = 1.44$.



Empirical CDF of spike size in log-log scale

Case study: EEX series



Comparaison of the real EEX series and the simulated series with estimated parameters. Which is which?