Sparsity Analysis of Energy Price Forecasting

Shi Chen
Wolfgang Karl Härdle
Brenda López Cabrera

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E.-Center for Applied Statistics and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de

EEX

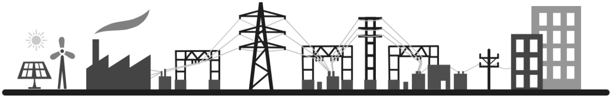
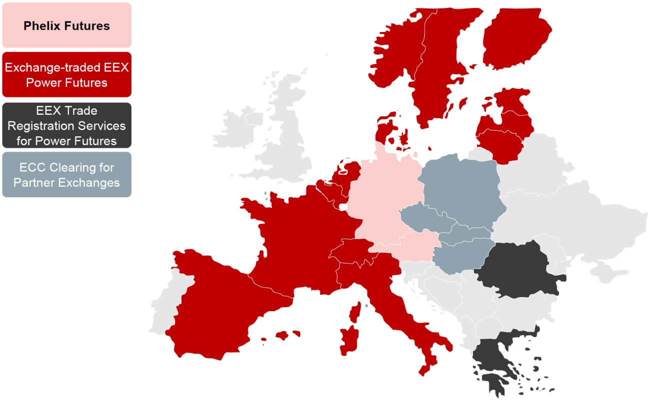


Figure 1: Source: Blue Star Energy

- European Energy eXchange develops, operates and connects secure, liquid and transparent markets for energy and related products.
- EEX provides the trading platform for the German/Austrian Power market.
- Phelix Futures: most liquid contract and benchmark for European power trading.

EEX Power Derivative

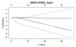


Source: EEX website

Phelix Futures

- Contract: A contract settling against the average power spot market prices of future delivery periods.

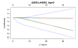
- Underlying: Physical Electricity Index determined daily by EPEX Spot Exchange.
  - Phelix Base: average price of the hours 1 to 24 for electricity traded on spot market.
  - Phelix Peak: average price of the hours 9 to 20 for electricity traded on spot market.



Phelix Futures

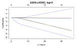
- Find price drivers and important variables in big system.
- Energy companies' decision-making mechanisms.
- Hedge against weather risk; Predict energy price.

Source: Home utilities blog



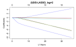
Motivation

- How all these prices of different contracts interact with each other?
- Which variables are crucial for the whole system?
- Model selection, variable selection.



Motivation - ctd

- Due to large number of variables in the system, some sparsity assumption must be imposed for the sake of an accurate estimate.
- Large dimension comes from,
  - varieties of future contracts
  - large lag in VAR model to avoid the correlation of error term



Outline

1. Motivation ✓
2. VAR model
3. Iterative Sure Independent Screening
4. Estimation Results
5. Forecasting
6. Conclusion

VAR model

- The VAR(p) model is constructed according to Lütkepohl (2005),

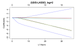
$$\begin{aligned} y_t &= \nu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t \\ &= \nu + (A_1, A_2, \ldots, A_p) \left(y_{t-1}^\top, y_{t-2}^\top, \ldots, y_{t-p}^\top \right)^\top + u_t (1) \end{aligned}$$

  - $y_t = (y_{1t}, y_{2t}, \ldots, y_{Kt})^\top$ is a $(K \times 1)$ random vector consisting K prices we have at time t, t from 1 to T.
  - ν is a $(K \times 1)$ vector of intercept terms.
  - $u_t = (u_{1t}, u_{2t}, \ldots, u_{Kt})^\top$ is a K-dimensional innovation process.

Assumption

- The coefficients $\nu, A_1, \ldots, A_p$ are assumed to be unknown in the following.
- y_t are partitioning into sample and presample values to facilitate the following analysis. Define,

$$
\begin{aligned}
Y &= (y_1, y_2, \ldots, y_T) \\
B &= (\nu, A_1, A_2, \ldots, A_p) \\
Z_t &= (1, y_t, y_{t-p+1})^\top \\
Z &= (Z_0, Z_1, \ldots, Z_{T-1})
\end{aligned}
$$



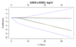
Compact form

- Hence for multivariate case, the model described in equation (1) can be rewritten as,

$$Y = BZ + U \tag{2}$$

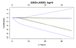
- The compact form (2) is equivalent to,

$$\text{vec}(Y) = (Z^\top \otimes I_K)\text{vec}(B) + \text{vec}(U) \tag{3}$$



Discussion

- If the vector ν is assumed to be zero, the total dimension of the model to be estimated is pK^2 and the total number of observations is KT.
- When $Kp > T$, we need lasso etc to estimate the model.
- Under normal assumption, the upper bound of error is positively correlated in $\frac{\log K^2 p}{T}$, part of oracle inequality.
- To improve the estimation, introduce the (iterative) sure independent screening.

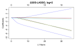


Basic Idea

- Concept of sure screening: Fan and Lv (2008)
  - Based on correlation learning which filters out the features that have weak correlation with the response.
- Let $\omega = (\omega_1, \ldots, \omega_p)^\top$ be a $p-$vector obtained by componentwise regression, i.e.,

$$\omega = X^\top y \tag{4}$$

  - y is n-vector of response.
  - X is $n \times p$ data matrix.

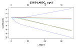


Basic Idea

- When there are more predictors than observations, LS is noisy.
- Consider ridge regression, let $\omega^\lambda = (\omega_1^\lambda, \ldots, \omega_p^\lambda)^\top$ be a $p-$vector obtained by ridge regression, i.e.,

$$\omega^\lambda = (X^\top X + \lambda I_p)^{-1} X^\top y \tag{5}$$

  - $\lambda > 0$ is a regularization parameter.
  - $\lambda \to 0$, $\omega^\lambda \to \hat{\beta}_{LS}$
  - $\lambda \to \infty$, $\lambda\omega^\lambda \to \omega$
- Componentwise regression is a specific case of ridge regression.

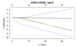


Penalized Estimators

- Let us start with consider model estimation and variable selection in a linear regression model,

$$y = X\beta + \varepsilon \tag{6}$$

  - $y = (y_1, \dots, y_n)^\top$ is an $n \times 1$ response vector,
  - $X = (x_1, \dots, x_p)$ is an $n \times p$ matrix with $x_j = (x_{1j}, \dots, x_{nj})^\top$, $j = 1, \dots, p$.
  - $\hat{\beta}$ denotes the coefficient estimator produced by the fitting procedure.
  - $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is an $n \times 1$ vector of iid random errors.



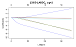
Penalized Estimators

- The lasso is a regularization technique for simultaneous estimation and variable selection. It estimates are defined as

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

$$= \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} x_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - λ is a tuning parameter.

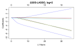


Penalized Estimators

- The lasso is a regularization technique for simultaneous estimation and variable selection. It estimates are defined as

$$\begin{aligned}\hat{\beta}_{LASSO} &= \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} x_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|\end{aligned} \tag{8}$$

  - λ is a tuning parameter.
  - The second term is ℓ_1-penalty, the coefficients shrinks toward 0 as λ increase.



Penalized Estimators

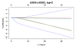
- Smoothly Clipped Absolute Deviation penalty - SCAD estimator is given by,

$$\hat{\beta}_{SCAD} = \begin{cases} \operatorname{sgn}(\omega)(|\omega| - \lambda)_+ & \text{when } |\omega| \leq 2\lambda \\ \dfrac{\{(a-1)\omega - \operatorname{sgn}(\omega)a\lambda\}}{a-2} & \text{when } 2\lambda < |\omega| \leq a\lambda \\ \omega & \text{when } |\omega| > a\lambda \end{cases} \qquad (9)$$

- The continuous differentiable penalty function for SCAD estimator is defined by,

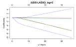
$$p_\lambda^{'}(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \qquad (10)$$

where $a > 2$ and $\beta > 0$



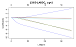
Iterative SIS - (I)SIS

1. Apply SIS for initial screening, reduce the dimensionality to a relative large scale d;
2. Apply a lower demensional model selection method (such as lasso, SCAD) to the sets of variables selected by SIS;
3. Apply SIS to the variables selected in the previous step;
4. Repeat step 2 and 3 until the set of selected variables do not decrease.



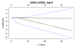
Data: Phelix Futures

- Product overview: EEX offers continuous trading data.
- Load profiles: base, peak and off peak
- Maturity:
  - Day/ Weekend Futures
  - Week Futures
  - Month Futures
  - Quarter Futures
  - Year Futures

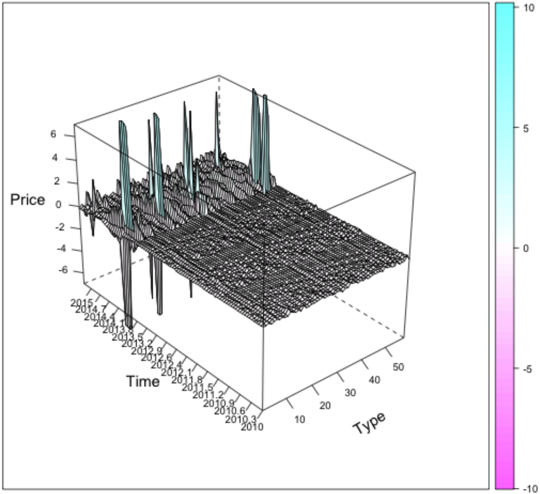


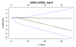
Data

- Recall Underlying: Physical Electricity Index determined daily by EPEX Spot Exchange.
  - Phelix Base: average price of the hours 1 to 24 for electricity traded on spot market.
  - Phelix Peak: average price of the hours 9 to 20 for electricity traded on spot market.
- Spot price:
  - Hours: $00-01, \dots, 23-24$
  - Blocks: Base Monthly, off-peak 01-08, off-peak 21-24, Peak Monthly

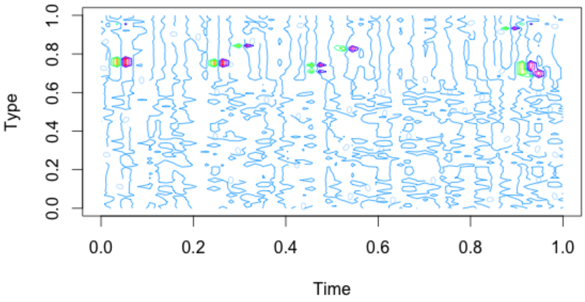


First-order Difference

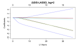




Data



- Time span: 30.09.2010 - 31.07.2015
- Type: 58 kinds of contracts



Estimated Coefficients

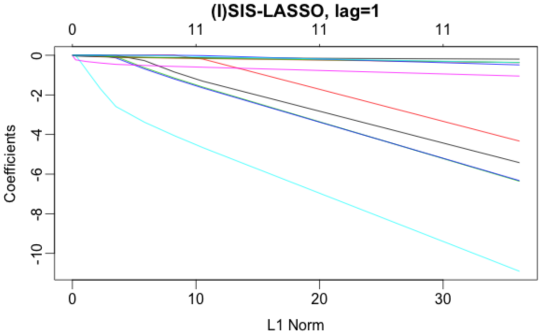
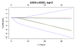


Figure 2: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

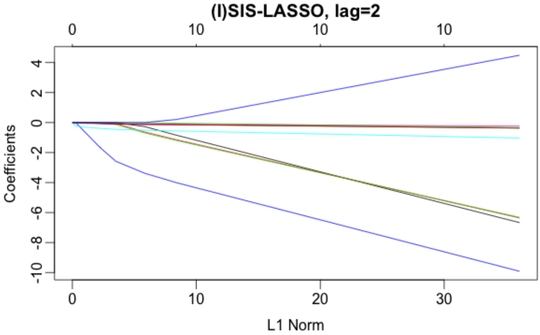
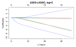


Figure 3: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

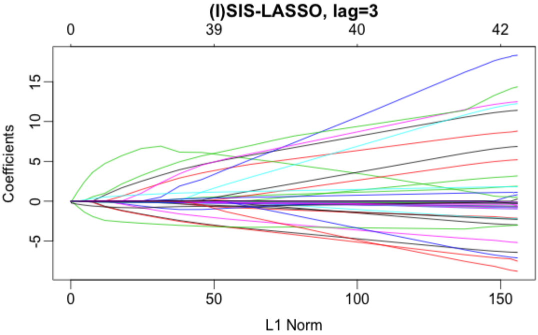
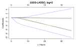


Figure 4: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

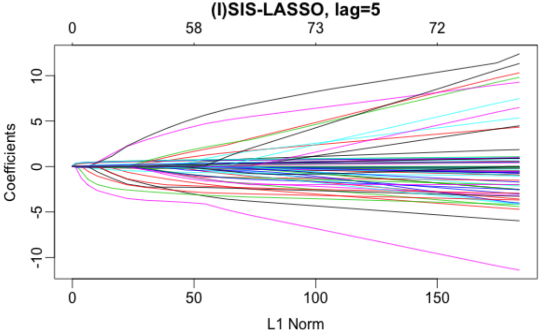
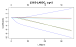


Figure 5: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

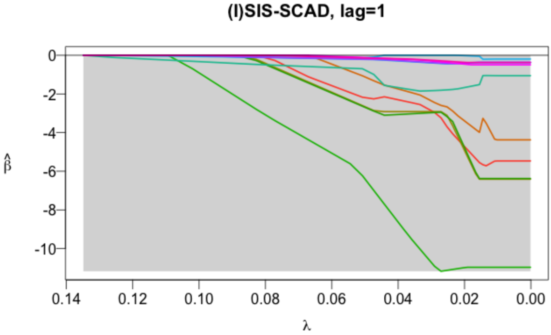
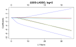


Figure 6: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

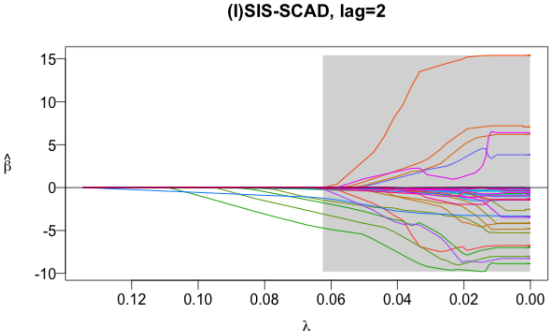
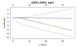


Figure 7: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Estimated Coefficients

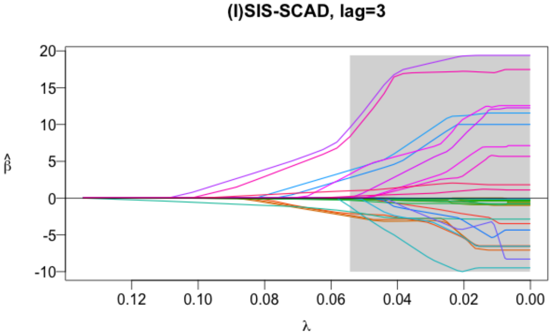
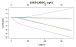
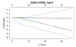


Figure 8: Each curve corresponds to a variable. It shows the path of its coefficient against the l1-norm of the whole coefficient vector at as λ varies.



Samples

- Pre-sample selected: 30.09.2010 - 28.11.2014
- Sample selected: 31.12.2014 - 31.07.2015
- High dimension due to,
    - varieties of future contracts
    - large lag in VAR model to avoid the correlation of error term

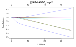


Lag Length Selection

- The lag length for the VAR(p) model may be determined using model selection criteria.
- General approach:
  - Fit the VAR(p) models with different lags $p = 0, \ldots, p_{max}$,
  - Choose the value of p which minimizes some model selection criteria.
- Model selection criteria for VAR(p),

$$IC(p) = \log|\hat{H}(p)| + \varphi(K, p)c_T \tag{11}$$

  - $\varphi(K, p)$ is a penalty function.
  - c_T a sequence indexed by the sample size T.



Lag Length Selection - ctd

- The residual covariance matrix without a degrees of freedom correction is defined as,

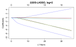
$$\hat{H}(p) = \frac{1}{T}\sum_{t=1}^{T} u_t^\top u_t$$

- The three most common information criteria are the Akaike (AIC), Schwarz-Bayesian (BIC) and Hannan-Quinn (HQ):

$$AIC = \log|\hat{H}(p)| + \frac{2}{T} pK^2 \quad (12)$$

$$HQ = \log|\hat{H}(p)| + \frac{2\log\log T}{T} pK^2 \quad (13)$$

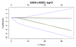
$$BIC = \log|\hat{H}(p)| + \frac{\log T}{T} pK^2 \quad (14)$$



Model Selection

Model	AIC	HQ(C)	BIC
(I)SIS-lasso, $p = 1$	4.5686	4.7249	5.7864
(I)SIS-lasso, $p = 2$	4.5006	4.6426	5.6076
(I)SIS-lasso, $p = 3$	7.7034	8.3143	12.4637
(I)SIS-lasso, $p = 5$	7.0839	8.1209	15.1652
(I)SIS-SCAD, $p = 1$	4.5714	4.7277	5.7892
(I)SIS-SCAD, $p = 2$	6.1043	6.1043	9.5782
(I)SIS-SCAD, $p = 3$	7.2559	7.6820	10.5770

Table 1: The three most common information criteria: the Akaike (AIC), Schwarz-Bayesian (BIC) and Hannan-Quinn (HQ) are compared.



Forecasting Errors

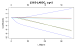
- Recall the VAR(p) model,

$$\begin{aligned} y_t &= \nu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t \\ &= \nu + (A_1, A_2, \ldots, A_p) \left(y_{t-1}^\top, y_{t-2}^\top, \ldots, y_{t-p}^\top \right)^\top + u_t \end{aligned}$$

- Pre-sample: 09.2010 - 11.2014, Sample: 12.2014 - 07.2015

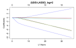
Lag	(I)SIS-lasso	(I)SIS-SCAD
$p = 1$	0.0697	0.0697
$p = 2$	0.0670	0.0701
$p = 3$	1.9598	0.1413
$p = 5$	0.1397	-

Table 2: MSE of forecasting during 31.12.2014 - 31.07.2015



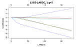
Results

- Model selected: VAR(2) with (I)SIS-lasso.
- Key contracts: LPXBHR15 Index, LPXBHR06 Index, LPXBHR07 Index,LPXBHR08 Index.
    - The spot price bid from 14-15h is the price driver.
    - The spot price bid from 14-15h, 05-06h, 06-07h, 07-08h are essential for the prices of the phelix futures.
    - Contracts interact with each other..



Conclusion

- Apply variable selection technique to electricity data analysis, for high-dimensional case.
- Find the core element in the big system.
- Further work: IR in the setting of sparsity.

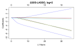


References

Fan, Jianqing, and Jinchi Lv.
Sure independence screening for ultrahigh dimensional feature space.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70.5 (2008): 849-911.

Fan, Jianqing, and Rui Song.
Sure independence screening in generalized linear models with NP-dimensionality.
The Annals of Statistics, 38.6 (2010): 3567-3604.

Lütkepohl, Helmut.
New introduction to multiple time series analysis.
Springer Science & Business Media, 2005.



References

Fan, Jianqing, and Runze Li.
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96.456 (2001): 1348-1360.

Tibshirani, Robert.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), (1996): 267-288.