Forward price simulation

Hao Wu and Adrian Dragulescu

November 10, 2008

1 Background

Long term price simulation is useful in many aspects. Simulated forward prices can be used in calculation of long term Value at Risk (VaR), potential exposure (PE), optimal asset allocation, etc.

In this project we simulated forward curves for all commodity futures contracts in portfolio. The simulation is done solely from the finance aspect, ie., forward price is simulated based on historical prices. We assumed the log prices follow OU processes and did simulation from there. Because of the strong correlation among curves, all curves should be simulated correlatedly. Given the scale of the data (30,000+ curves) it's computationally infeasible to simulate all the data together. Principal component analysis (PCA) was applied to reduce the dimension of the data. Futhermore, we built a hierarchy among the curves so that the simulation can be done sequentially. Simulation results look good and the performance is acceptable.

2 Data exploration

Understanding the historical data is crucial in forward simulation. We want to look at the data from many different angles and try to comprehend the relationships in them. This is not an easy task given the scale of the data. This section will give some detailed description of what I have learned from the data.

2.1 Natural Gas

Being a major part of the portfolio, there are over 400 natural gas curves. Henry Hub gas price is the standard for North America natural gas market so we will start from it. Figure

?? shows historical prices of all Henry Hub future contracts for past 200 days, with each line representing a future contract. The top one plots in the original scale and the bottom figure plots prices centered around 0. We can see at the first glance that most of the curves are all very similar except for a few. From the bottom figure one can see that the prices for near futures have increased a lot while the prices for further futures stayed flat. As a result the curves crossed at around 130 days ago. This reflects the traders' view on natural gas market. They think the natural gas price will increase in the near future but drop back to normal after some time. Our simulation should reflect this trend.

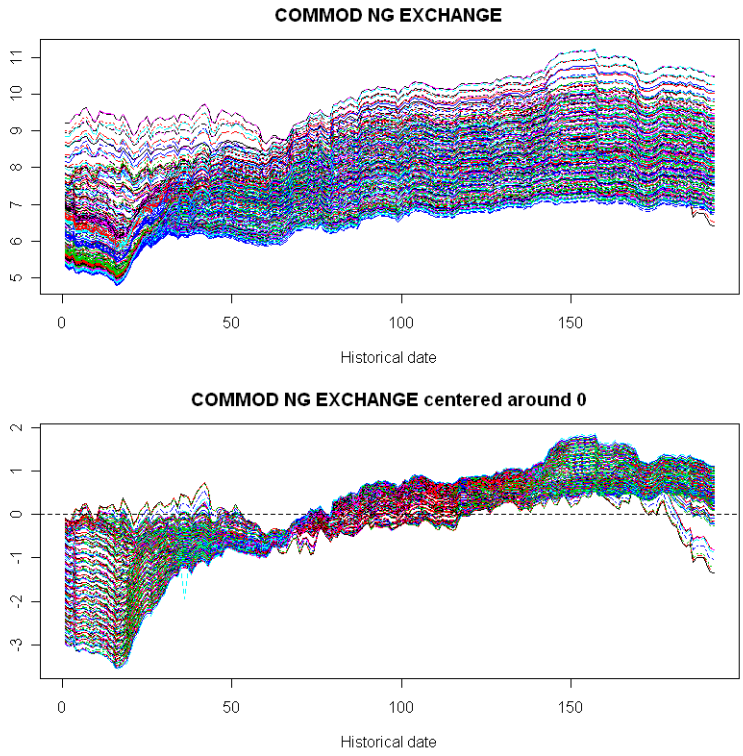

Figure 1: Historical prices for Henry Hub gas future.

Figure ?? shows the same plots, but for January and August contracts only. The curves behaved similarly so the patterns are consistent for all contracts.

Now look at the Henry Hub gas prices from another angle. Figure ?? plots the historical Henry Hub prices versus contract months. Each line represents a pricing day. The striking sinusodal pattern reflects the seasonality and our simulation must capture this. The wide spread in the tail reflects people's uncertainty of distant future. People orginally thought the gas price will increase in the future but changed their mind recently. Note that this change of future perspective depends on many factors such as new drilling, storage and

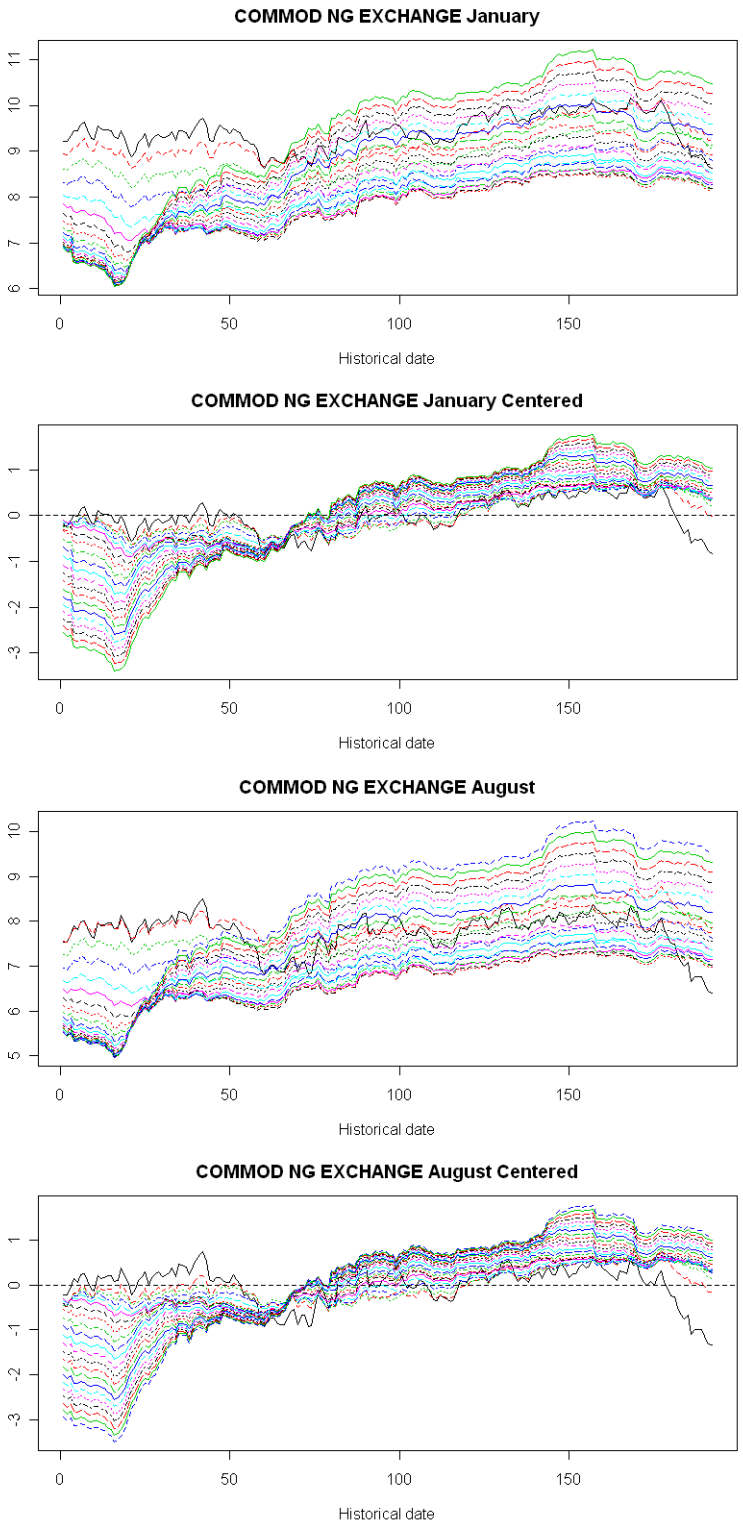

Figure 2: Historical prices for Henry Hub gas future - January and August contracts only.

pipeline construction, economic growth, etc. It is very difficult to simulation solely based on historical prices. We can see later that our simulation result shows very little of that, which suggests a more complex model might be necessary.

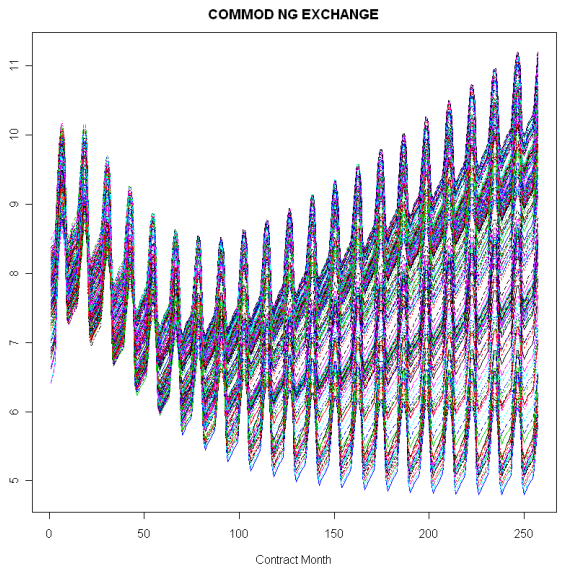

Figure 3: Henry Hub historical prices versus contract month.

Figure **??** shows the same plot for January and August contracts. By taking the contract for a specific month we removed the seasonality. The up/down patterns are preserved.

Since we have observed from Figure **??** that prices for all contract months are similary, we want to quantify the similarity by looking at their correlations. Figure **??** shows the heat map for historical price correlations among future contracts. Darker means less correlation. We can clearly see the correlations are above 0.9 for contracts in nearby months. The correlation dropped to around 0.7 if two contracts are far apart, say, 10 years. All correlations are above 0.9 for contracts 10 years later.

Now we want to look at curves at different locations. According to natural gas traders the whole North America market can be divided into 17 regions. Each region has a leading curve. In figure **??**, the left panel shows the historical prices of August 2007 contract for those 17 leading curves. We can see after centering all curves are almost the same except for the one from Rockie area. The right panel shows the correlation among those prices.

Figure **??** shows the hierarchical cluster for regional curves. The result shows that the markets physically close were clustered together, which makes sense.

To summarize the findings in natural gas historical prices, we have:

1. For one curve, prices for all future contracts moved similarly.

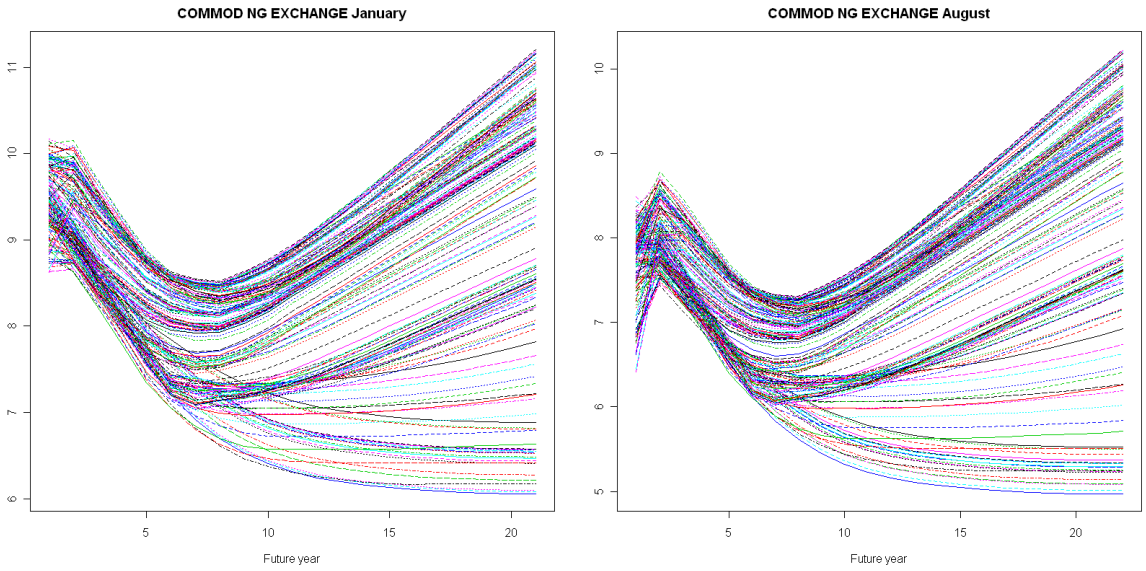

Figure 4: Henry Hub historical prices versus contract month - January and August contracts only.

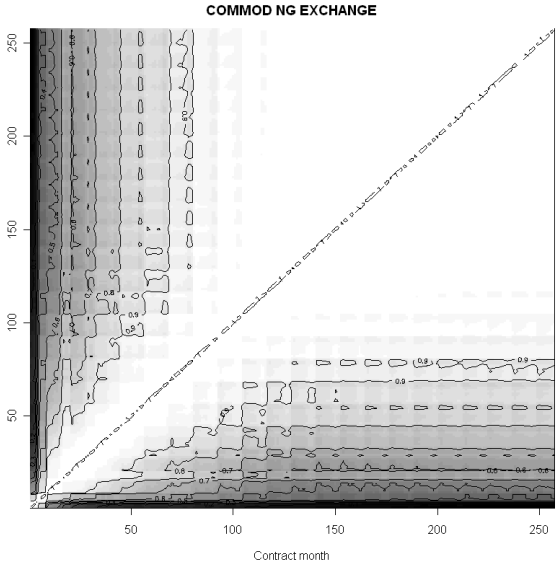

Figure 5: Henry Hub historical prices correlation among future contracts.

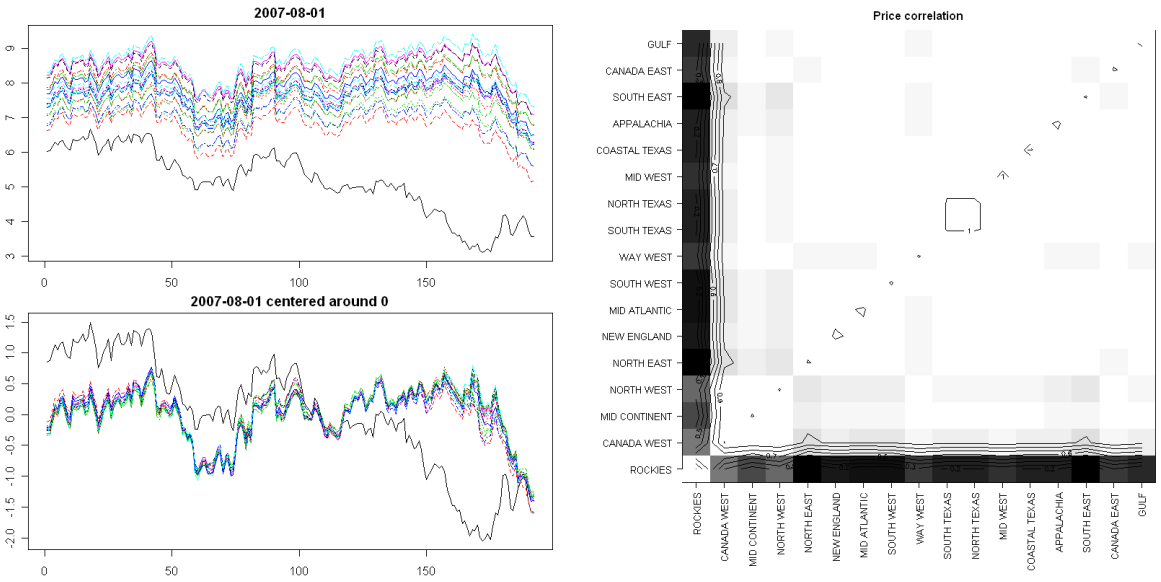

Figure 6: Historical prices of August 2007 contracts for 17 regional leading curves. Left: Historical prices; Right: correlation among historical prices.

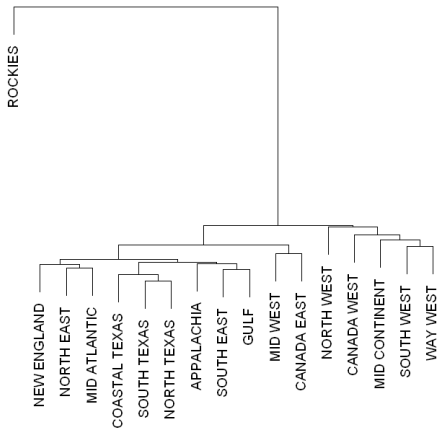

Figure 7: Hierarchical cluster on historical prices of August 2007 contract for 17 regional leading curves.

2. Future contracts show seasonality, with the contracts for January and August the most expensive.

3. For one contract month, all curves are very similar except for those in Rockie Mountain area.

2.2 Coal

There are over 90 domestic coal curves. Figure **??** plots the historical prices of all future contracts for COMMOD COL EXCHANGE, which is the standard for domestic coal market. From the left panal one can see that prices for all contracts are highly correlated, with the curves crossed a little bit. The right panel shows the prices by contract month. Unlike the natural gas prices, there is no seasonality. The prices increase along with contract month, which means people think the coal price will increase in long term.

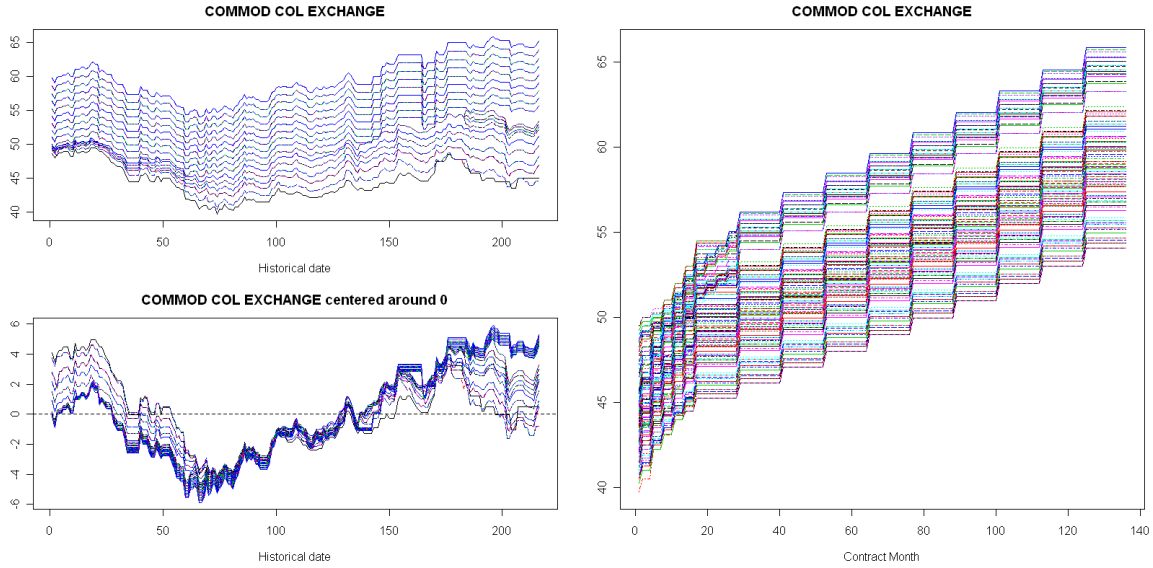

Figure 8: Historical prices for COMMOD COL EXCHANGE. Left: prices for all future contracts. Right: Prices versus contract months.

2.3 Crude Oil

There are only 8 Oil curves (WTI). Figure **??** plots the historical prices of all future contracts for COMMOD WTI EXCHANGE, the standard for crude oil market. Again from the left panal one can see that prices for all contracts are highly correlated. The right panel shows

the prices by contract month. There's no seasonality and the prices are almost flat along with future months, which means people currently think the long term crude oil price won't change much.

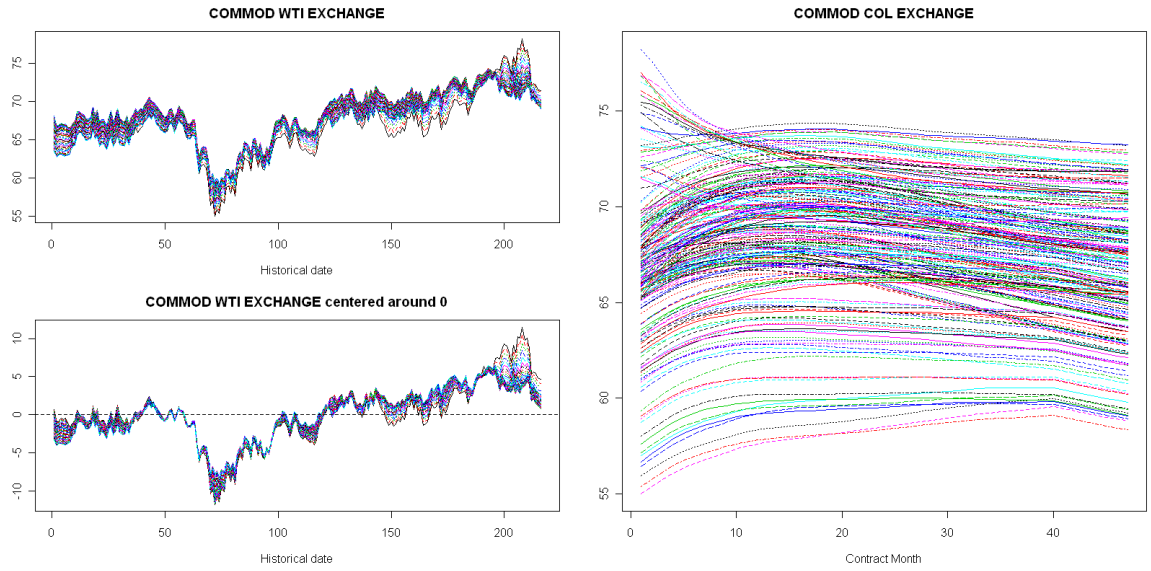

Figure 9: Historical prices for COMMOD WTI EXCHANGE. Left: prices for all future contracts. Right: Prices versus contract months.

2.4 All fuels

We want to understand the relationship among different types of fuels. We took three fuel exchange curves, shifted and rescaled the historical prices to make them have the same mean and standard deviation. Figure **??** shows the three transformed prices for for September 2007 contract. One can see these prices are mildly correlated with correlation coefficient being around 0.5.

2.5 Electricity

There are around 20 regional electricity markets in North America and part of Europe. Each market could have as many as 400+ curves (such as PWY), or as little as less than 10 curves (such as PWF). Totally there are close to 2000 curves so the scale of the data is huge.

Figure **??** shows the historical prices of all future contracts for one of the major PWY peak hour curves. Left panel shows the daily price changes and right panel shows the prices

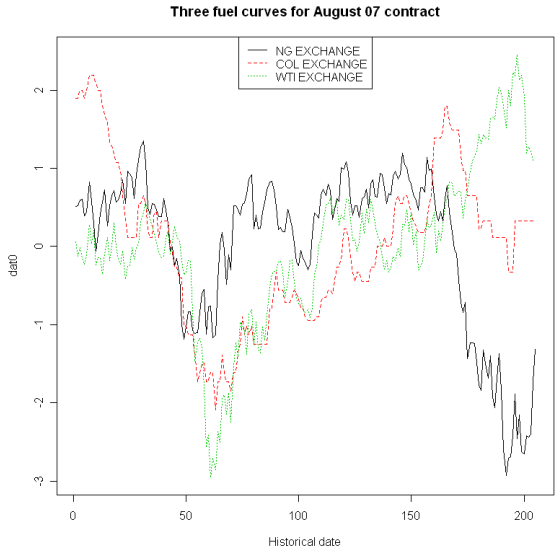

Figure 10: Historical prices (shifted and scaled) of three major fuel curves.

versus contract months. We can see that (1) the prices for different contract are still very similar, although the correlation is lower than those in natural gas; and (2) Electricity prices show seasonalities with the prices in summer higher than the prices in winter (note that month 0 is August).

Figure **??** shows all PWY curves for the same contract month. They are very similar but the similarity is not as strong as in natural gas data.

In order to find the group structure in electricity prices, We did K-means clustering on PWY data. The result shows that data can be roughly clustered into 2 or 3 groups. 5X16 curves show great similarity and always grouped together, so are the 7X8 curves. 2X16 curves floats between these two groups, depending on market location and contract month. I guess that that the grouping depends on many other factors such as demand, fuel prices, etc. The grouping is not very important in forward curve simulation so in order to simplify it, we divide the electricity curves into peak and offpeak groups and put 2X16 into the peak groups. Figure **??** shows a K-mean clustering result of 2 groups for PWY September 2007 conract. The figure at top are roughly the offpeak curves and the one at bottom are peak curves.

Now we want to explore the relationship between electricity and fuel prices. Figure **??** plots the historical prices of major peak and offpeak curves for September 2007 contract, versus three types of fuels. One can see that the peak curve (left panel) is tightly correlated with the NG curve and has little correlation with other two fuels, whereas the offpeak curve

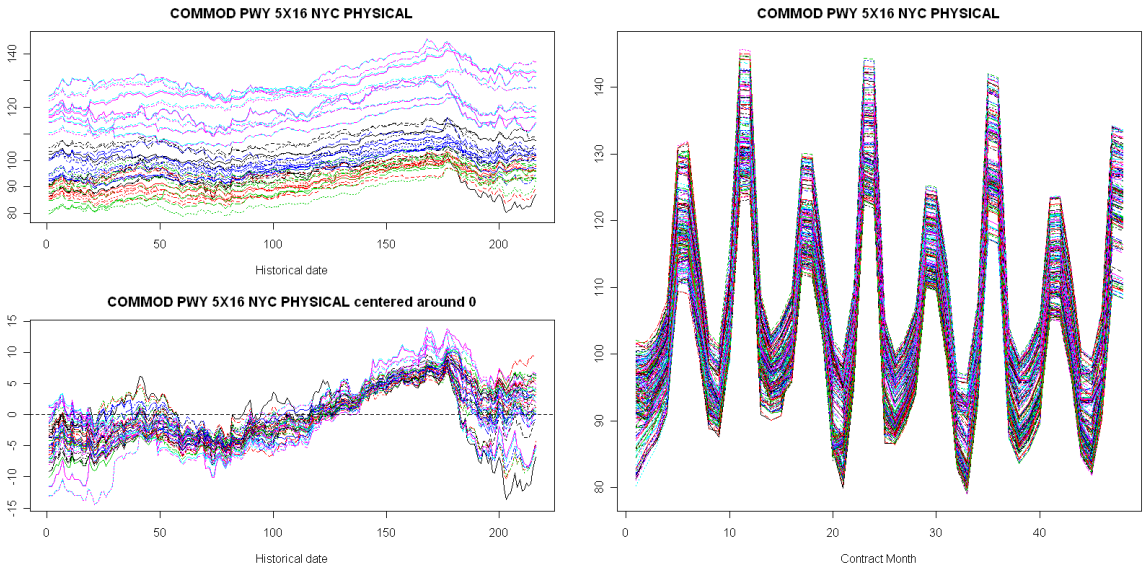

Figure 11: Historical prices for COMMOD PWY 5X16 NYC PHYSICAL. Left: prices for all future contracts. Right: Prices versus contract months.

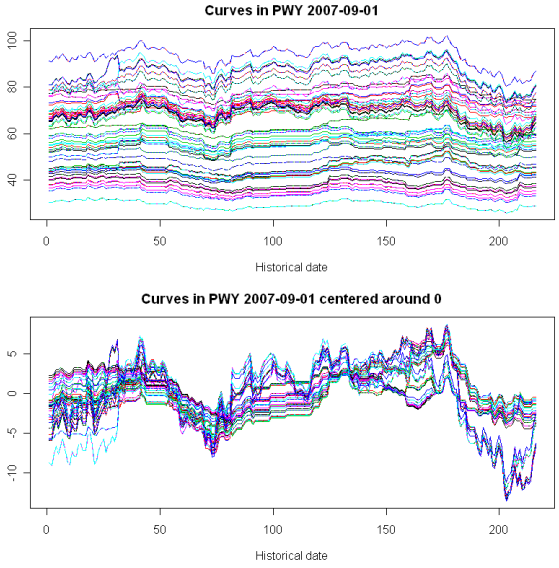

Figure 12: Historical prices for all curves in PWY, September 2007 contract.

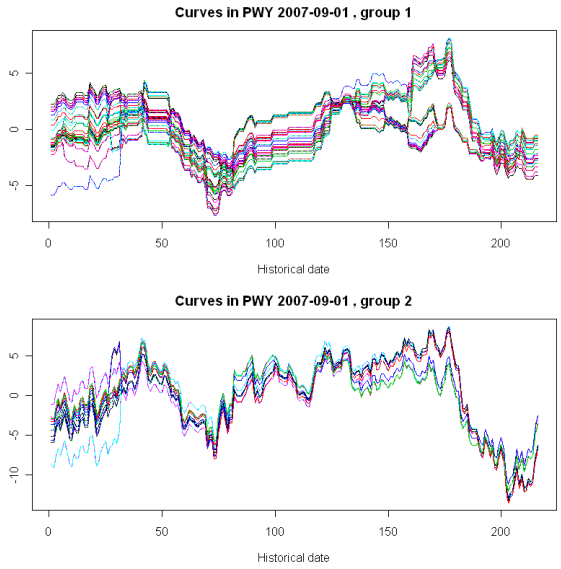

Figure 13: K-mean clusters of PWY historical prices for September 2007 conract.

(right panel) mildly correlates NG and COL curves. This makes perfect sense because the quick started but expensive gas turbines were usually served as peakers. As a result the marginal cost of electricity during peak hours are proportional to gas prices. During offpeak hours the cheaper coal buring generators serve most of the base load so the marginal cost of electricity correlated more with coal prices.

Since the correlation between electricity and fuels depends on load, they might show seasonalities. Figure **??** shows the correlation between electricity and fuels by contract month. One can see for peak curve the correlation to NG is always high, above 0.9 most of the time. The correlatino to COL and WTI were low in the begining but climbed up quickly. For offpeak curve the correlation to NG and COL both started from around 0.6, and the corrlation to WTI started low. The correlations to all three fuels show some seasonality with the correlation are lower in early winter.

2.6 Others

There are many other markets including freight, emission. I did spend much time on them so the description on those curves will be skipped. Since the scale of those data is not very big, I will simulate them all at once, market by market.

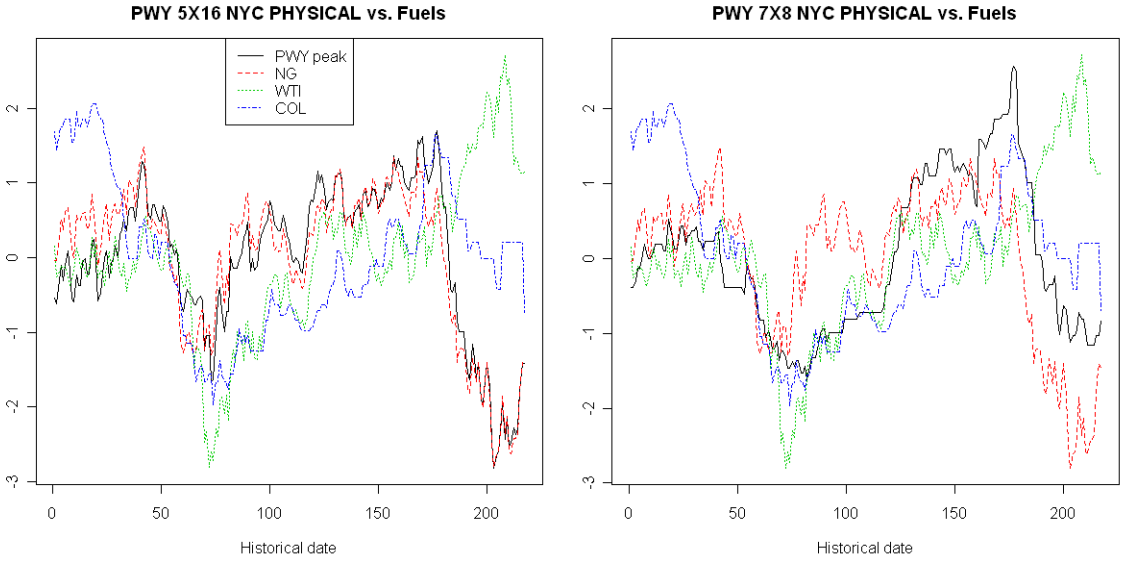

Figure 14: Major PWY peak and offpeak curves vs. Fuels, September 2007 contract.

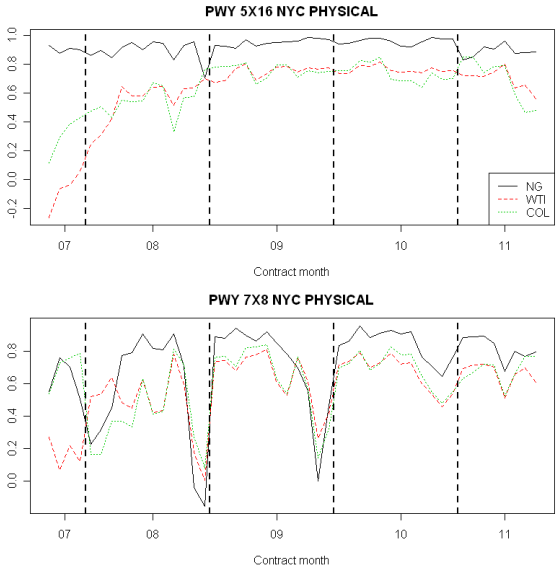

Figure 15: Correlation between major PWY peak and offpeak curves vs. Fuels, for contract months over next four years.

2.7 Volatility

We suspected that the volatility will increase when a contract approaching the maturity date. So we studied the volatility of some curves. Figure **??** shows the daily log returns of four expired contracts for NG EXCHANGE. It's not obvious that the volatility changes change a long with time.

Figures **??** and **??** show the same plots for one offpeak and one peak curves for PWY. Interestingly the volatilities seem to increase when approaching maturity for January and July contracts, but not so for April and October contracts.

We did the same plots for two PWJ curve, one peak and one offpeak, as shown in Figures **??** and **??**. We observed a similar pattern as in PWY curves. This is a interesting finding, which might suggest the volatility changes show some seasonality. Maybe during peak seasons there are many unpredictable factors and the traders made wrong decision more often. We will have to explore more curves to make a reasonable claim. Modeling volatility is difficult and one have to make some parametric assumptions. Throughout this work I made an assumption that the volatilities didn't change along with time.

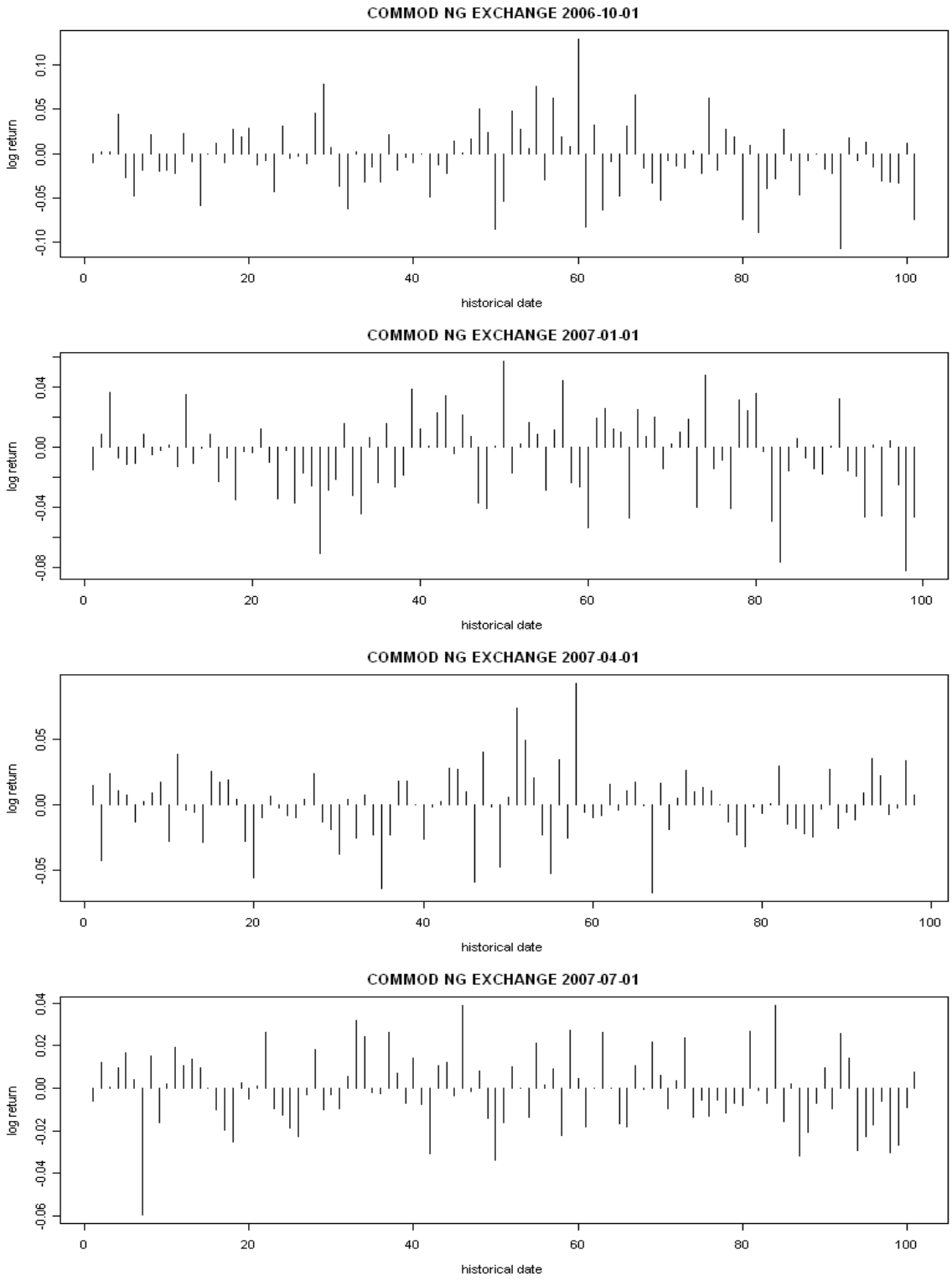

Figure 16: Daily log returns of four expired contracts for NG EXCHANGE

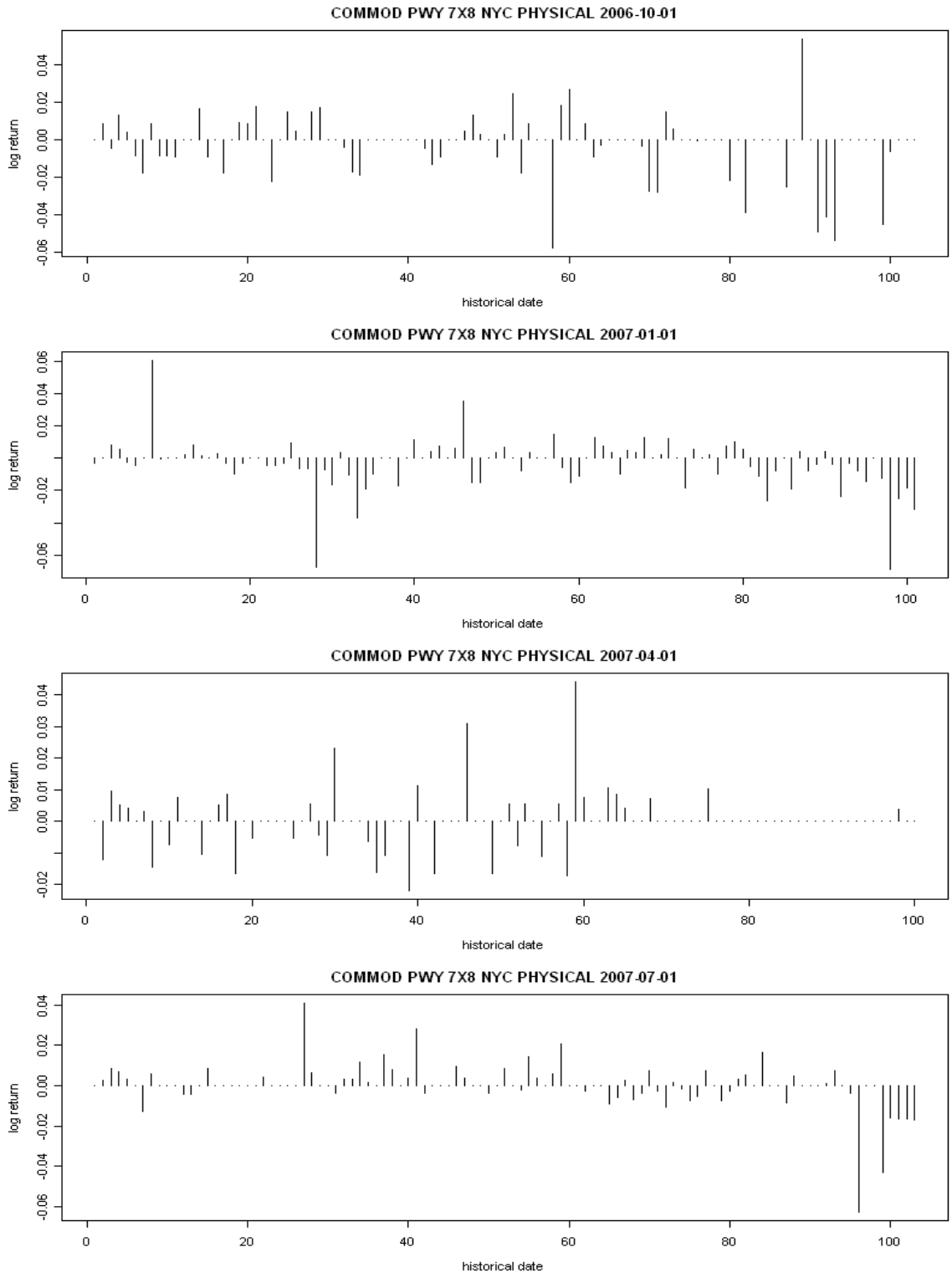

Figure 17: Daily log returns of four expired contracts for PWY 7X8 NYC PHYSICAL

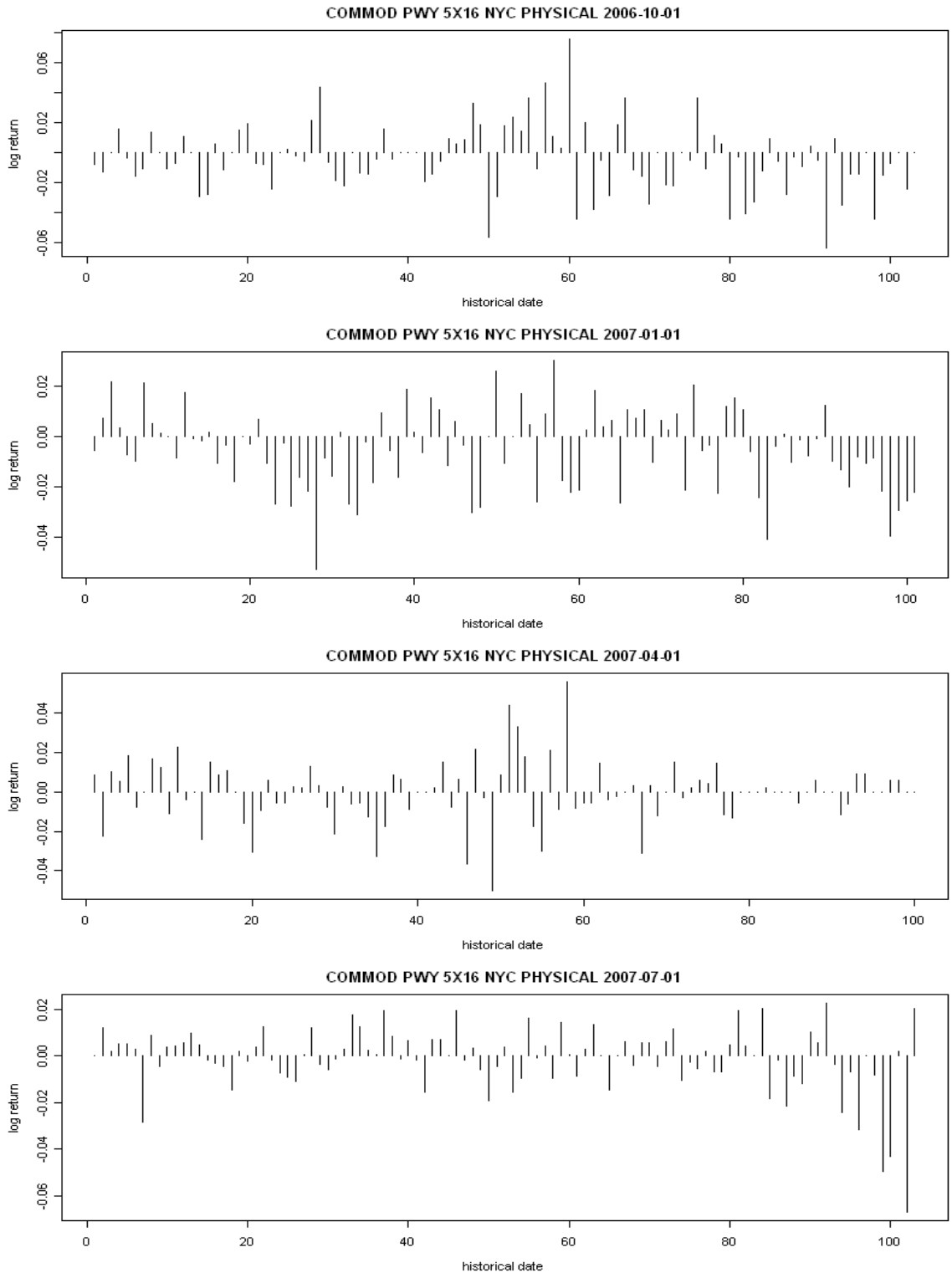

Figure 18: Daily log returns of four expired contracts for PWY 5X16 NYC PHYSICAL

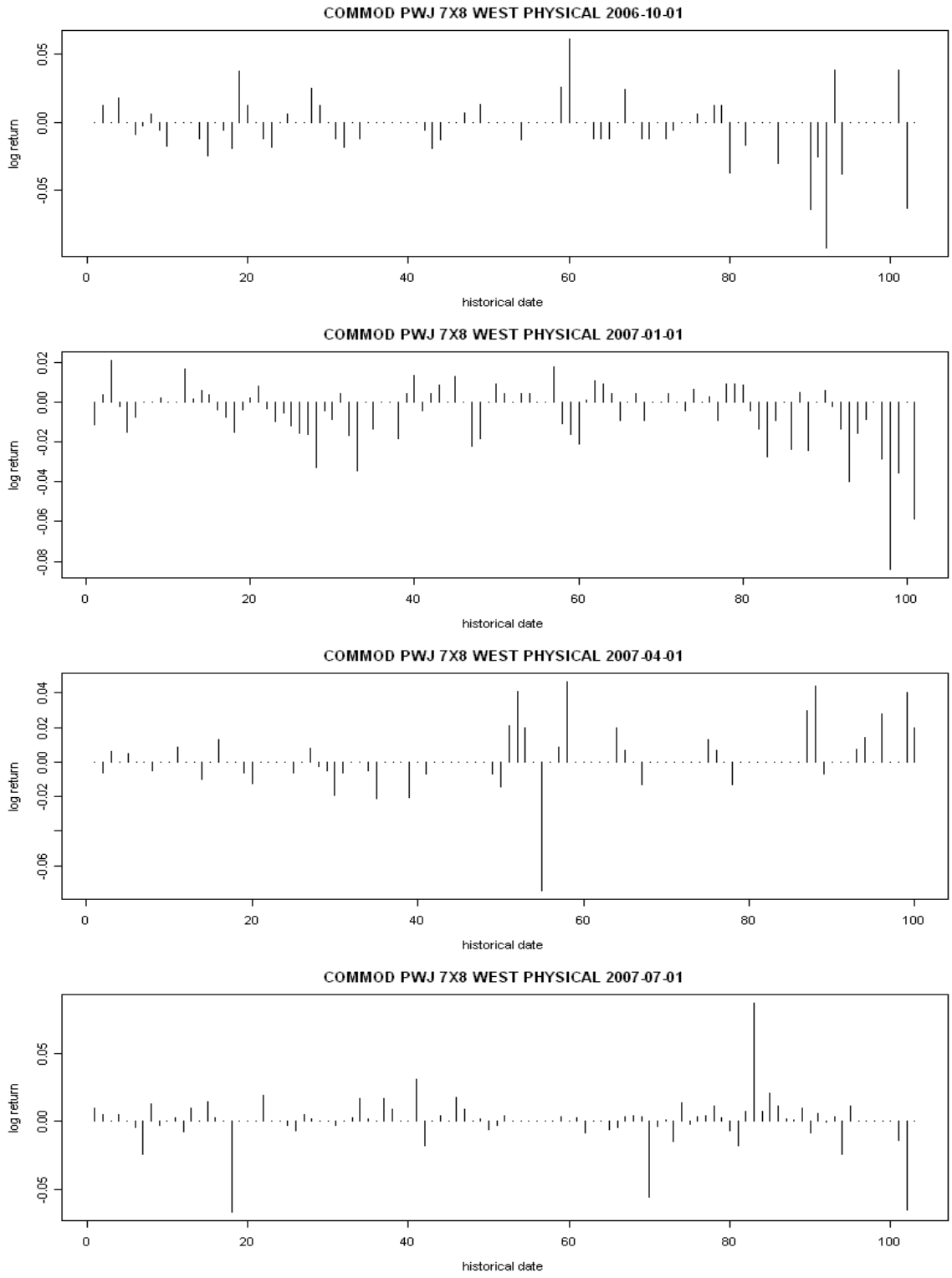

Figure 19: Daily log returns of four expired contracts for PWJ 7X8 DOMHUM PHYS

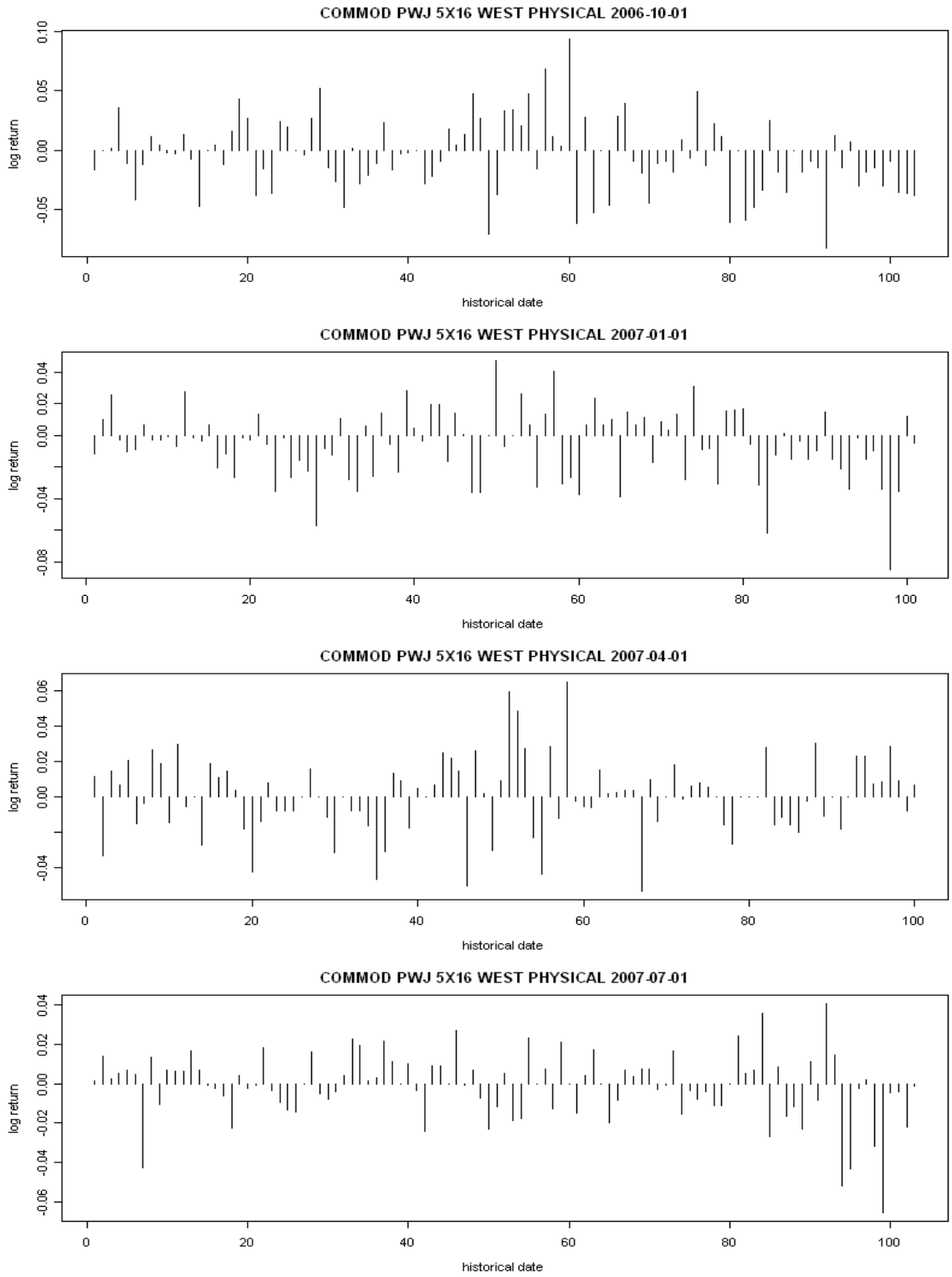

Figure 20: Daily log returns of four expired contracts for PWJ 5X16 DOMHUM PHYS

3 Method

This section will give some technical details for the methods used in this project.

3.1 Orthogonalization and dimension reduction

There are over 600 curves in portfolio, each one has contracts for 30 to 120 future months. So totally there are over 30,000 curve-month combinations. As illustrated in previous section, in each market there are strong correlations among curve-months in two direction: for each curve the prices for all contract months are highly correlated. Also for each contract month, all curves are correlated. So ideally all curve-months should be simulated together. That requires to handle a 30,000 by 30,000 correlation matrix, which is not feasible computationally.

Principal component analysis (PCA) is a widely used technique in data dimension reduction. It extracts common features from correlated data (usually of high dimension) and summarize them by a few independent data sets called Principal component (PC). The dimension can be greatly reduced without lossing much information. Probably more important than dimension reduction in this problem is that the resulting PCs are independent so that we can break the correlations at one direction and do simulation on each group of PCs separately.

More specifically, for each curve we applied PCA for its log historical price of all contract months and kept the first K PCs. Doing so broke the correlations among contract months. The correlation among curves were posessed in the PCs now, e.g., PC1's for all curves are highly correlated, so are the PC2's and so on. PC's are orthogonal so simulation can be done for each PC independently.

To make it formal, let column vector $X^c_{t_m} = [x^c_{t_m}(1), x^c_{t_m}(2), \ldots, x^c_{t_m}(T)]^T$, where $x^c_{t_m}(t)$ denotes the log price at (historical) time t of future contract with delivery at month t_m for curve c, with $c = 1, \ldots, C$, $m = 1, \ldots, M$ and $t = 1, \ldots, T$. The correlations among $X^c_{t_m}$ are in two directions:

$$cov([X^c_{t_1}, X^c_{t_2}, \ldots, X^c_{t_M}]) = \Sigma^c \text{ for } \forall c, \text{ and } cov([X^1_{t_m}, X^2_{t_m}, \ldots, X^C_{t_m}]) = \Omega_m \text{ for } \forall m.$$

We can do PCA at either direction but we have observed that the correlations among contract months are higher and probably more consistent. So it makes more sense to do PCA for each curve. For commodity c, let $\mathbf{X^c} = [X^c_{t_1}, X^c_{t_2}, \ldots, X^c_{t_M}]$ be a T by M matrix, we can find a rotation matrix W^c of dimension M by K and let $\mathbf{Y}^c = \mathbf{X}^c W^c$. $\mathbf{Y}^c$ is a T by K matrix with independent columns, Columns of $\mathbf{Y}^c$ are first K PCs for commodity c. Let Y^c_k,

$k = 1, \ldots, K$ be columns of $\mathbf{Y}^c$, we now have

$$cov([Y_1^c, Y_2^c, \ldots, Y_K^c]) = \Lambda^c, \forall c, \text{ and } cov([Y_k^1, Y_k^2, \ldots, Y_k^C]) = \Omega_k^*, \forall k.$$

Λ^c are diagonal matrices so PCA transformed the correlated "month" domain to independent "PC" domain. Correlation among curves are possessed in Ω_k^*, which can be estimated from data. In simulation, for each k in $1, \ldots, K$, we can independently genearte forward k^{th} PC (correlatedly among curves) then transform them back to get the forward price in original scale.

Due to the great similarities among curves, the first a few PCs usually can explain most of the variances. Below table shows the percentage of variance explained by the first 10 PCs for a PWY peak curve. One can see that cumulatively the first 5 PCs explains almost 99% of the total variances.

	PC 1	PC 2	PC 3	PC 4	PC 5
% var explained	0.7962	0.1240	0.0424	0.0168	0.0083
Cumulative % var explained	0.7962	0.9202	0.9625	0.9794	0.9877
	PC 6	PC 7	PC 8	PC 9	PC 10
% var explained	0.0032	0.0018	0.0012	0.0010	0.0007
Cumulative % var explained	0.9909	0.9928	0.9939	0.9949	0.9956

Figure **??** shows the plot for the first 10 PCs. The scales of the curves dropped quickly.

3.2 Curve pedigree

After breaking the correlation among months by PCA, it seems we are ready to simulate. But for each PC there are still over 600 curves. A 600 by 600 covariance matrix is manageable but inefficient. More importantly, in the future if we want to adopt more complicated methods such as econometrics model, it will be difficult to implement. So we are looking to further reduce the dimensionality of the problem. From Data section we can see curves are more similar when their markets are physically closer. So it's a natural thinking to divide the curves into groups by market. Based on these ideas we built a dependency structure for curves based on their commodity type and location.

We first pick one commodity reference curve for each type of fuels (NG, COL, FCO, WTI). These four reference curves consists of the top level of the pedigree. Under that for natural gas, we divide the North America market into 17 regions. Each region has a regional

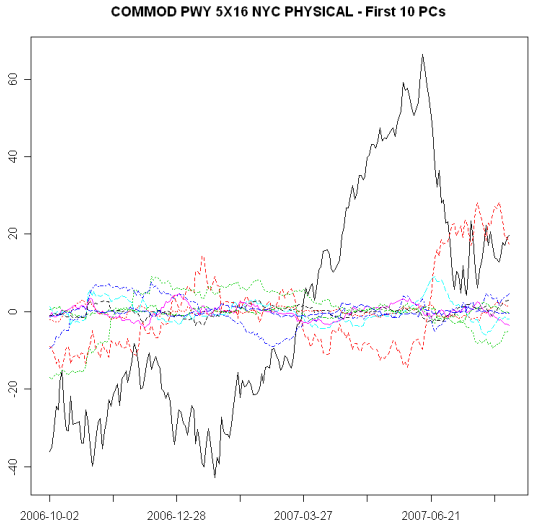

Figure 21: First 10 PCs for a PWY peak curve

reference curve. These regional reference NG curves are children of the NG commodity reference curve. The rest of the NG curves are children of their regional reference curve respectively. For COL, FCO and WTI, we didn't divide the market by region at this time so all of the curves are direct children of their commodity reference. It will be easy to add another layer for regions for those commodity in the future.

There are about 20 electricity markets by location. A market can have as many as 400 curves (such as PWY). From Data section we see all these curves can be roughly put into two groups, peak and offpeak. So we pick two reference curves in each electricity market, one for peak and one for offpeak. We set the parents of these electricity market reference curves to be three types of fuels: COL, WTI and its nearby NG regioanl refernce curves. For example the parents for PWY reference curves are WTI EXCHANGE, COL NYMEX PHYSICAL and NG TRAZN6 NY PHYSICAL. It is possible that some parents are not correlated with the children curve. So in that sense the parents we assigned are just "potential parents". A model selection procedure (discussed in later section) will be used to pick the real parents.

We didn't put any structure on the rest of the commodities (freight, emission, etc.) at this time. It would be easy to assign dependency in the future.

An illustration of the curve pedigree is shown in Figure **??**.

The curve pedigree is saved as a flat table in an Excel file, each curve in a row, with around 2400 rows in total.

Now we can do simulations based on the curve pedigree. We first simulated parents

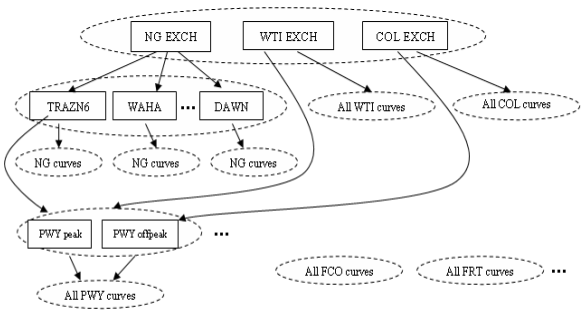

Figure 22: Curve pedigree

curves then simulate the children curves conditional on parent curves.

3.3 Simulation of parent curves

Parent curves are defined as those without a parent. They don't have to have children. For example all CO2 curves are deemed as parent curves because we didn't assign any parent for them. Parent curves are simulated in groups in order to keep the correlations among them. Four commodity refernce curves for fuels are in the same group. In each of the "other markets", all curves are in the same group.

We made a critical assumption here that each commodity price follows a OU process. With the consideration of correlations, all curves in the same group will follow correlated OU processes. To make it formal, assume the log price for commodity c, contract month t_m follows:

$$dX^c_{t_m}(t) = -\gamma^c_{t_m}\{X^c_{t_m}(t) - \theta^c_{t_m}\}dt + \sigma^c_{t_m}dW^c_{t_m}(t),\ m = 1, \ldots, M,\ c = 1, \ldots, C$$

The Wiener processes $W^c_{t_m}(t)$ for $\forall m, c$ are correlated.

Prior to this step, we had applied PCA on log historical prices and broke the correlation among contract months. Since the PCs are linear transformation of log prices, they will follow OU processes too. The PCs follow:

$$dY^c_k(t) = -\gamma^{*c}_k(Y^c_k(t) - \theta^{*c}_k)dt + \sigma^{*c}_k dW^{*c}_k(t),\ k = 1, \ldots, K,\ c = 1, \ldots, C$$

PCA ensures that there are no correlation among different PCs for the same commodity, ie, $cor(W^{*c}_i(t), W^{*c}_j(t)) = 0$ for $\forall i \neq j$. But for the k^{th} PC of all curves, the Wiener processes $W^{*c}_k(t), c = 1, \ldots, C$ are still highly correlated. I will call the same PC (e.g., PC1) for all

curves "PC group" thereafter. It is worth to mention that there could be correlations for different PCs between different commodity, ie, $cor(W_i^{*c_1}(t), W_j^{*c_2}(t)) > 0$ for some $i \neq j$. There is no theoretical result to justify the indepdendence among them. But in practice I found very little correlation, mainly because of the strong correlation within each PC group. If, say, PC1 and PC2 are independent for commodity 1, and PC1's for commodity 1 and 2 are very similar. It is safe to claim that PC1 of commodity 2 and PC2 of commodity 1 would be (almost) independent.

To simulate forward prices, we only need to simulate each PC group independently, then transform them back. Within each PC group, all the curves (PC) are still correlated. In order to simulate correlated OU processes, we did another round of PCA (PCA on PCs) to orthogonalize them. The results (PCs of the PCs) are independent OU processes and can be simulated easily.

To summarize, the steps for simulating parents curves are:

1. Do PCA on each curve to break the correlation among contract months,
2. Iterate following substeps for all PC groups to simulate forward PCs correlatedly:
   (a) Do PCA on the PCs in the same group to break the correlation among curves. This gives independent OU processes.
   (b) For each process from previous step, estimate OU parameters and simulate forward curves.
   (c) Transform back to get forward PCs for each PC group.
3. Gather the forward PCs for all PC groups then transform back to get the forward log prices.

One VERY important thing is that sometimes the price didn't follow OU process. For example is the prices kept going up for all historical days, the estimated γ will be negative, which means there is greater force to push price to continue go up when it is further away from the long term mean (θ). At this time what I did was to make γ a very small positive number (10^{-5}) if the estimated value is negative. Of course this will produce wrong result. However in practice this was only found in higher order PCs, which contributes little in price decomposition. So the simulation result still looks fine. There are some possible solutions to this. One is to use longer historical time. Another more complicated way is to decompose the time horizon into "up", "down" and "flat" regions. In each region we can remove the longer term trend and fit OU process on residuals. Anyway this remains an open question and is an area for future reseach.

3.4 Simulation of children curves

After having all parent curves simulated, we are in place to simulate the children curves. The children curves should be generated by group to keep their correlation. Note that a group of children curves could have different parents. For example, all electricity market reference curves are in the same group because we want to correlate different market, yet their natural gas parents are different. We assume the children curves prices follow the parent curves prices. A multiple linear regression will be the easiest way to model that dependency.

Using a set of new notations, let $y(t)$ be log historical prices for a child curve, and $x_p(t)$ be log historical prices for its p^{th} parent curve, $p = 1, 2, \ldots, P$. The most straightforward idea is to regress the children's daily prices on the parents' daily prices, e.g.,

$$y(t) = \beta_0 + \sum_{p=1}^{P} \beta_p x_p(t) + \epsilon(t). \tag{1}$$

This gives pretty good fit with R^2 above 0.85 for most of the curves. However there are some problems. First is that the regression could be "spurious" as referred by some literature so R^2 is not reliable. Secondly this model assumes that the children curve track parent curves tightly in a daily basis, which isn't necessarily true. The most important problem in practice is that there are strong auto-correlation among the residuals $\epsilon(t)$. In simulation we want to generate forward prices sequentially and we want the regression residuals are independent from day to day. Auto-correlated residuals are undesirable.

We then tried to regress the children log returns on parents' log returns, e.g.,

$$\Delta y(t) = \beta_0 + \sum_{p=1}^{P} \beta_p \Delta x_p(t) + \epsilon(t). \tag{2}$$

This model fits the data poorly with R^2 dropped dramatically to around 0.4 for many curves. The reason is obvious: the children and parent curves are cointegrated. Although the prices evolve together, their daily log returns could be very different.

Model **??** can be rewritten as:

$$y(t) = \beta_0 + y(t-1) + \sum_{p=1}^{P} \beta_p \{x_p(t) - x_p(t-1)\} + \epsilon(t).$$

It actually regress child's today's price $y(t)$ on its previous day's price $y(t-1)$ and parents' today and previous days' prices $x_p(t)$ and $x_p(t-1)$. The model impose constraints on the coefficients to force the coefficient for $y(t-1)$ is 1 and the coefficients for $x_p(t)$ and $x_p(t-1)$

are the same in magnitude but with opposite signs. We relieved these constraints and came up with the following model:

$$y(t) = \beta_0 + \beta_1 y(t-1) + \sum_{p=1}^{P} \{\beta_{p1} x_p(t) + \beta_{p2} x_p(t-1)\} + \epsilon(t). \tag{3}$$

Model **??** fits the data very well and there're very little auto correlation left in the residuals. This model can be related to model **??** as the following. Since there are auto-correlations in $\epsilon(t)$ in **??**, we can regress $\epsilon(t)$ on $\epsilon(t-1)$. The residuals from that regression has very little auto-correlation left. This residual regression is equivalent to model **??**.

Model **??** can also be related to Error Correction Model (ECM) developed for modeling cointegrated processes. From model **??** if we calculate the difference between $y(t)$ and $y(t-1)$ we get:

$$\Delta y(t) = \beta_1 \Delta y(t-1) + \sum_{p=1}^{P} \{\beta_{p1} \Delta x_p(t) + \beta_{p2} \Delta x_p(t-1)\} + \Delta\epsilon(t).$$

In ECM there's no $\Delta x_p(t)$ term because it assumes $y(t)$ and $x(t)$ co-evolved. Since we now have the luxury to know $x(t)$ beforehand, inclusion of $\Delta x_p(t)$ can improve the model fitting. Note that from ECM one cannot get model **??**. So our model is more liberal.

In practice, all $x(t)$ and $y(t)$ are PCA scores. They were centered around 0 so the estimated β_0 should be 0. Therefore we dropped the intercept from model **??** to remove possible numerical error and have our final model as:

$$y(t) = \beta_1 y(t-1) + \sum_{p=1}^{P} \{\beta_{p1} x_p(t) + \beta_{p2} x_p(t-1)\} + \epsilon(t). \tag{4}$$

Note that the parents we assigned to children curves are just "candidate parent". They don't have to be correlated with their children curves or the correlation could be different for different contract months. So we did parent curve selection by regression model selection. In model **??** if the effects for $x_p(t)$ and $x_p(t-1)$ are not significant (with p-value bigger than 0.05) parent p will be dropped from the model. Doing so gives the most parsimonious model and remove unnecessary noises.

Now consider multiple correlated children curves $y_i(t)$, $i = 1, \ldots, C$. The correlation among children curves is the combination of the correlation among their parents and the correlation in the residual effects. The first component was correctly simulated in parent curve simulation. The second component needs some attention. The correlations in residuals from model **??** underestimates the truth. Because the children curves are also cointegrated but for computational reason we didn't take that into account (there's no $\Delta y_2(t-1)$ in

regression model for $y_1(t)$). We tried and found that using residual correlation from model **??** gives poor result. The correlations in simulated forward prices are underestimated, from above 0.9 in historical prices to around 0.7 in forward prices. So we use the correlation in the historical prices as the correlation for residuals. This will artificially inflate the correlation but considering the loss of correlation in dimension reduction, this can serve as a compensation. The bad thing is that it forces the daily residuals for all curves to be very similar. However since the regression R^2 is usually very high (≥ 0.95) this effect is minor. Simulation shows very good results

Another aspect need to be mention is that our forward time steps are not constant. We simulated daily data for some days then do monthly. Should the residual $\epsilon(t)$ in model **??** be from the same distribution? The answer is yes. $\epsilon(t)$ can be (kind of) viewed as the cointegration process of the child and parent curves, so it should be stationary. Another point is that although we didn't impose OU assumption on children curves, they are almost OU since they follow their parents (which are OU) closely.

4 Result

This section will present some of the simulation results. Simulation settings are:

- Use 300 historical days to predict future. Exclude the weekends and holidays there are around 210 historical pricing days.
- Use 5 principal components to represent both the curves and the months. This will capture over 95% of the variance in data most of the time.
- Simulate daily until the end of next month, then do monthly for another 46 months. Totally there are around 80 forward time points.
- Simulation was done for contracts of 48 future months.
- Generate 1000 independent simulations.

The simulation can be done in a single PC (P4 2.8G with 2G RAM) within about 2 hours.

4.1 Simulation result for a single curve

Figure **??** shows the simulation results of NG EXCHANGE for four 2008 futures contracts. The fisrt half of the figures plots the historical prices. The second half is a heat map

representation of all the simulation for daily forward prices, where region with darker blue means more simulated forward prices fall in there. Five randomly chosen forward curves were plotted on top of the heat map. OU process assumption guarantees the variances of forward prices become constant after some time.

Figure **??** shows the violin plot for simulated monthly forward prices of NG EXCHANGE for four future contracts. One can see again the variance didn't blow up and the simulated forward prices are approximately normal.

Figure **??** shows one simulation for all NG EXCHANGE contract months. First panel shows forward prices, the fisrt half is for daily and second half is monthly. We can see the curves move almost parallely, so that the correlations among curve were preserved. Volatilities in monthly forward is much highly than daily forward, as it should be. Second panel plots the data from another angle, price versus contract month. The seasonality was well preserved. Bottom two panels plot the same thing for historical data. We can see the similarity between forward and historical prices.

Table **??** summarize the daily, weekly and biweekly volatilities of log prices for simulated NG EXCHANGE prices in log scale, compared with historical volatilities. I only showed that for 5 contracts. One can see the daily and weekly volatilities are faily accurate. Move to a bigger time horizon, the biweekly volatilities were underestimated. This result veries by curves (some were overestimated) but simulated long term volatilities are not very accurate in many cases. This suggests for those curves the OU process assumption might be incorrect or unstable(the estimation of parameters are noisy). Some other stochastic process assumptions are worth trying.

Table 1: Historical vs. simulated volatilities for NG EXCHANGE.

	2007-10-01	2007-11-01	2007-12-01	2008-01-01	2008-02-01	
Daily	Hist	0.0228	0.0185	0.0154	0.0144	0.0142
	Sim	0.0245	0.0194	0.0154	0.0142	0.0139
Weekly	Hist	0.0423	0.0347	0.0291	0.0275	0.0271
	Sim	0.0519	0.0390	0.0303	0.0273	0.0267
Biweekly	Hist	0.0557	0.0452	0.0378	0.0362	0.0358
	Sim	0.0409	0.0356	0.0284	0.0258	0.0254

Figure **??** shows the simulated vs. historical heat content for PWY 5X16 NYC PHYSICAL. One can see the distribution of simulated values are consistent with history.

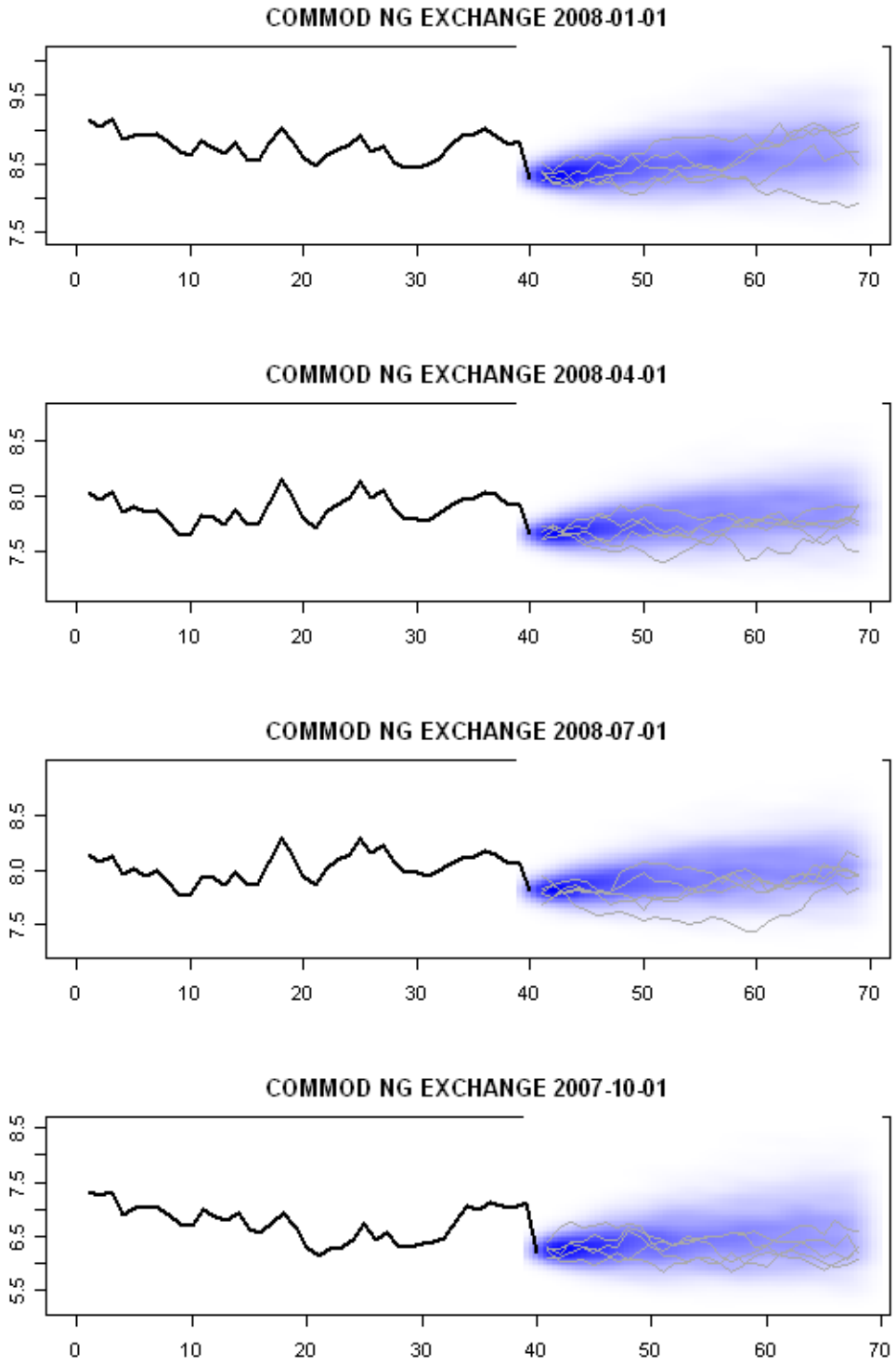

Figure 23: Simulation results for NG EXCHANGE: daily forward prices.

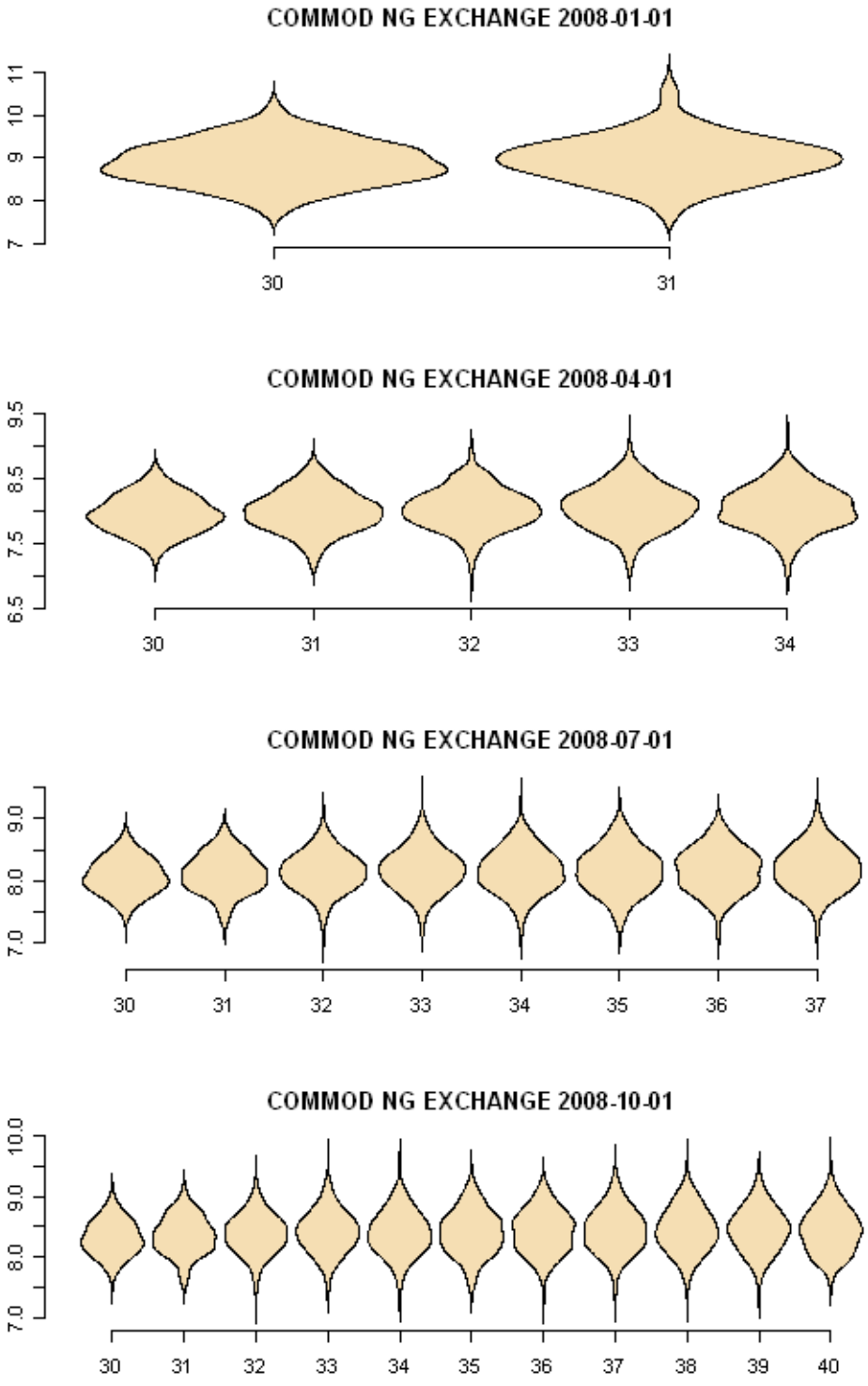

Figure 24: Simulation results for NG EXCHANGE: violin plot for monthly forward prices.

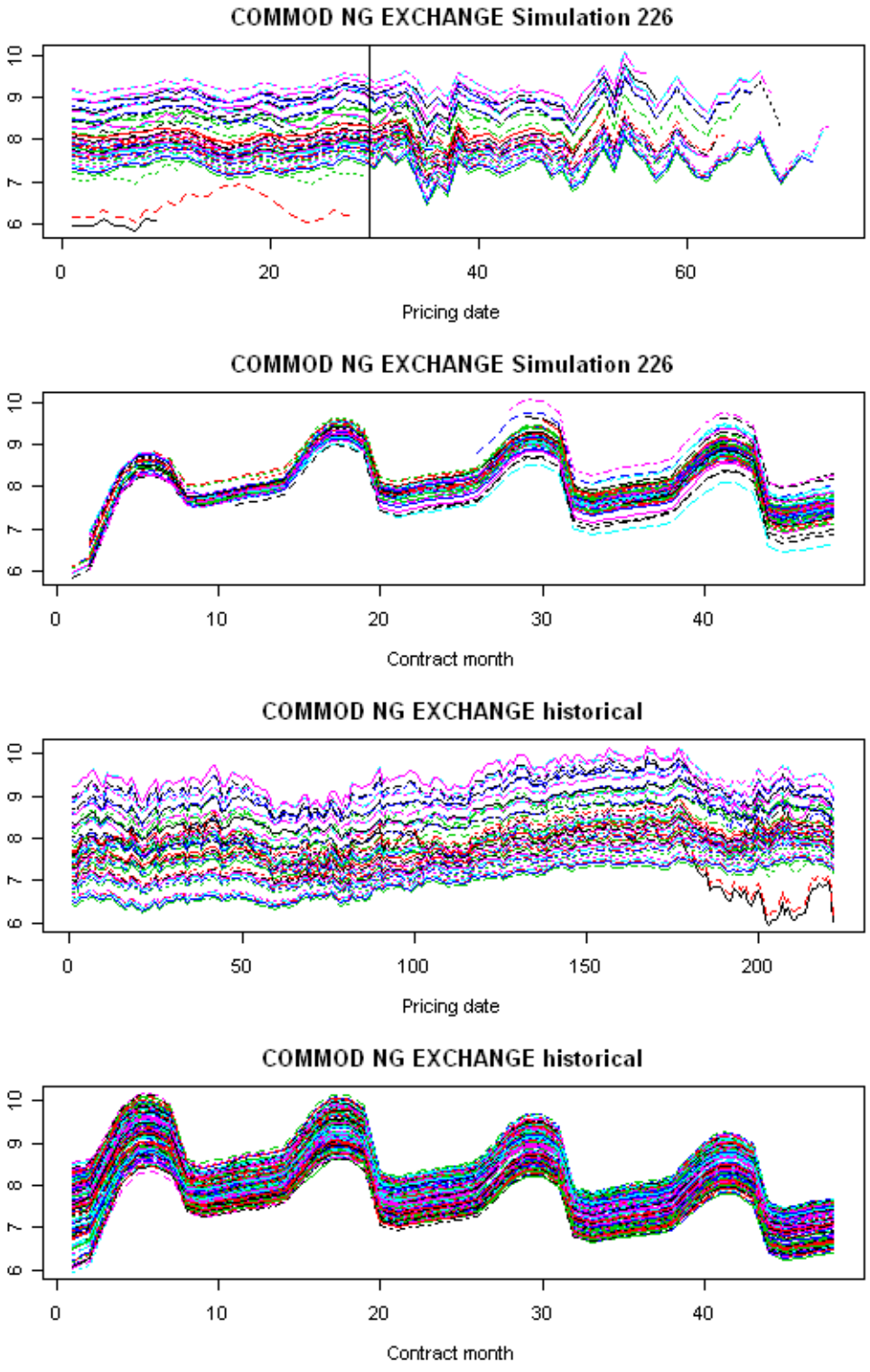

Figure 25: One Simulation for NG EXCHANGE all contract months, versus historical data.

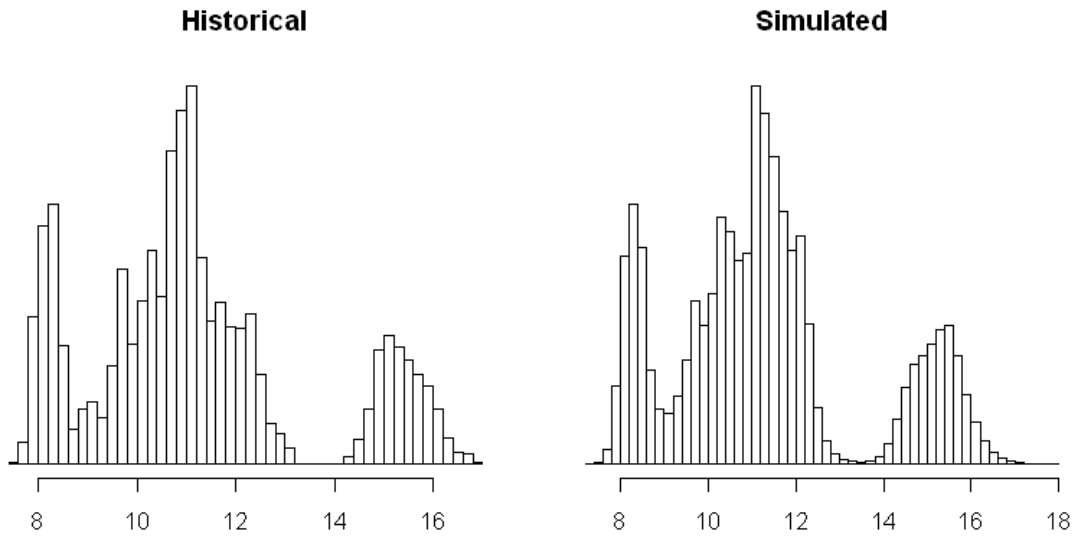

Figure 26: Heat content for PWY 5X16 NYC PHYSICAL.

4.2 Multiple curves

Now we move to view the simulation results for multiple curves. We want to make sure that the correlations among curves were preserved in the simulated prices.

Figure **??** shows one randomly chosen simulation for three important NG curves: NG TENZN6 PHYSICAL, NG TRAZN6 NY PHYSICAL, and NG DOMSP PHYSICAL. These are children curves of NG EXCHANGE so they were simulated together in the same group. We can see historically these curves are highly correlated. Table **??** shows the average simulated price correlation vs. historical. The correlations are a little bit smaller than historical but the error is tolerable.

Table 2: Historical vs. simulated correlations among three NG curves.

	Historical			Simulated		
	TENZN6	TRAZN6 NY	DOMSP	TENZN6	TRAZN6 NY	DOMSP
TENZN6	1.0000	0.9988	0.9983	1.0000	0.9737	0.9640
TRAZN6 NY	0.9988	1.0000	0.9976	0.9737	1.0000	0.9763
DOMSP	0.9983	0.9976	1.0000	0.9640	0.9763	1.0000

Now we move to children curves in different groups. We picked three electricity curves

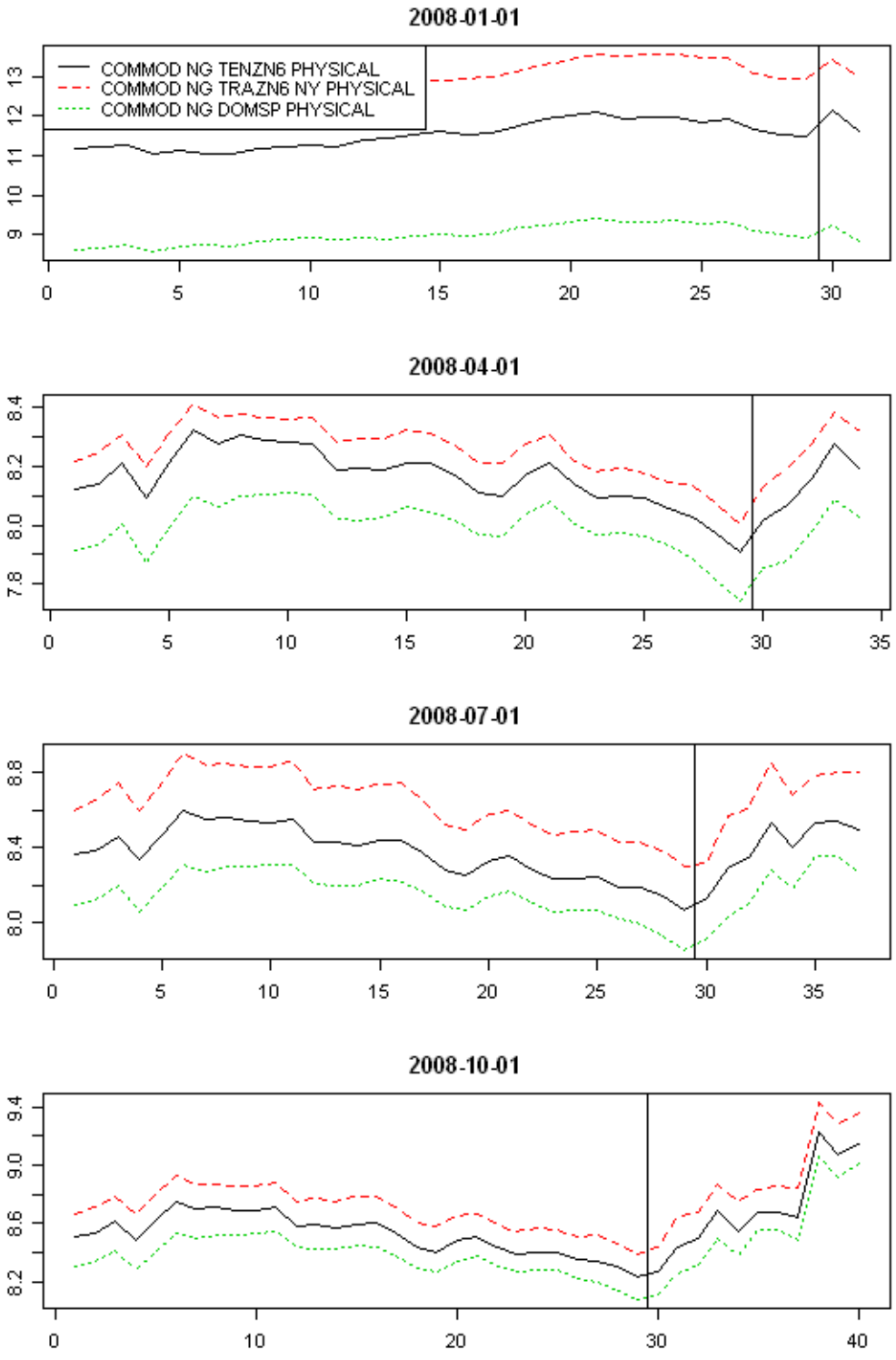

Figure 27: One simulation for three Natural Gas curves.

PWX 5X16 BOS PHYSICAL, PWJ 5X16 BGE PHYSICAL and PWY 5X16 NYC PHYSICAL. Those curves belong to different markets so correlation among them are passed on from their parents. Figure **??** plots result from one simulation. The prices was shifted and scaled so that they have the same mean and standard deviation.

Table **??** summarizes the correlation among them. This time the correlations in simulation were obviously underestimated, dropped to around 0.8 from 0.9.

Table 3: Historical vs. simulated correlations among three electricity curves.

	Historical			Simulated		
	5X16 BOS	5X16 BGE	5X16 NYC	5X16 BOS	5X16 BGE	5X16 NYC
5X16 BOS	1.0000	0.8673	0.9317	1.0000	0.8052	0.8205
5X16 BGE	0.8673	1.0000	0.9414	0.8052	1.0000	0.8813
5X16 NYC	0.9317	0.9414	1.0000	0.8205	0.8813	1.0000

A little calculation can show why we got this result. Let X, Y and Z be random variables. $cor(X, Y) = \rho_1$, $cor(Y, Z) = \rho_2$. Let the correlation between X and Z be ρ. When ρ_1 and rho_2 are faily large (bigger than $\sqrt{(}2)/2$, to be exact), ρ can be as low as $cos(cos^{-1}(\rho_1) + cos^{-1}(\rho_2))$. When $\rho_1 = \rho_2 = 0.99$, we have $\rho \geq 0.96$, which can be seen from Table **??**. For $\rho_1 = \rho_2 = 0.9$, $\rho \geq 0.62$. We can see when ρ_1 and rho_2 dropped a little bit, ρ dropped quickly. So when the correlations between children and parents are not VERY high (0.9 is high correlation but not high enough), correlations in simulated children could drop very quickly. One solution for this problem is to further cut the regions into smaller subregion, and assign more parent curves to the children. That way the children can correlate each other through many different paths and the result will be more reliable.

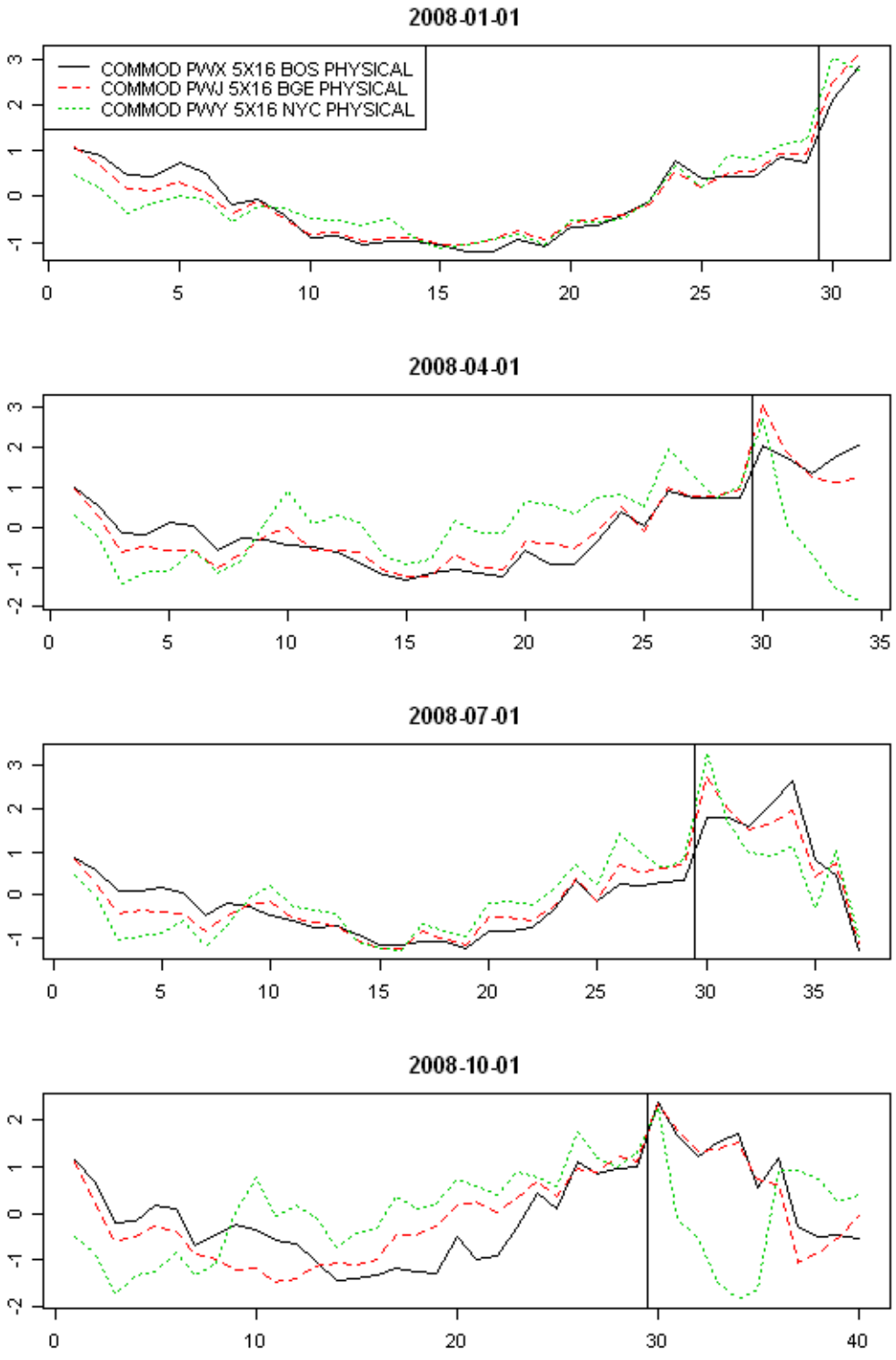

Figure 28: Simulation for three electricy curves.

5 Conclusion and future works

In this report we have proposed an algorithm for forward prices simulation. Given the large scale of the data we applied principal component analysis to reduce the dimensions and orthogonalize the curves. One essential assumption is that the prices are following OU processes. Parameters for the processes were estimated from historical data and forward curves were generated using the same parameters. A hierarchy for curves were built so that we can first simulate the parent curves, then simulate the children curves accordingly. Simulation results make good sense. The correlation structure and volatilities were well preserved.

The program was imlemented using open source programming language `R`. Computation was done parrally using open source software package `Condor`.

There are many areas can be worked on to improve the simulation. The major problems include:

1. How to deal with the cases when the assumption of OU process is incorrect. As discussed before we can either use a longer historical time or use Hidden Markov Model (HMM) type of approach to divide the time period into "up", "down" and "flat" periods. Then the trend can be removed and OU assumption can be applied on the residuals. The transition and emission probabilites can be estimated empirically and used in forward simulation. This might work well for short term but I would guess the long term simulation results will be similar because the trend will cancel in long term. Also this will introduce much more computations.

2. Regression models of children curve on parent curves. Currently a linear model is implemented to characterize the dependency but we could choose a different set of basis to capture the possible nonlinearility in the data.

3. Instead of using PCA to do dimension reduction, one can use nonlinear dimension reduction method, such as Locally Linear Embedding (LLE) . As described by Chen *et al.* the nonlinearility in electricity curves can be well captured by a few (like 3) intrinsic component, instead of many (currently 10) PCs. However according to what I observed, there are the nonlinearility is not obvious even in electricity curves. As a result the first 5 PCs can explain over 95% of the total variance so the gain from using LLE is not great. Plus the computation of LLE and restruction requires many other extra steps, such as finding nearest neighbors, constrainted least square, etc., so the overall computation might be even more. It seems to be that some simplified version of

LLE is possibe and promising. Maybe we can do one round of LLE on all curve/month combinations instead of two rounds of PCA and reduce some computation.

4. In the simulation we assume the volatilities didn't change along with time. From what we have observed in the data that is not true sometimes, especially in the electricity curves. There are several ways to model the volatility as a function of time to maturity. ARCH/GARCH type of model is an obvious choice. Or we can put a functional form on the volatility and estimate the paramters from historial data.

5. Finally we took complicated (and hopefully more accurate) approaches by introducing other factors, such as weather forcast, economy growth, transmission/storeage, etc. This should be done at the parent curves level.