Sparse Manifolds Graphical Modelling with Missing Values: An Application to the Commodity Futures Market

Loann David Denis Desboulets

Sparse Manifolds Graphical Modelling with Missing Values: An Application to the Commodity Futures Market

Loann D. DESBOULETS[†]

[†] CNRS, EHESS, Centrale Marseille, AMSE, Aix-Marseille University,
5-9 Boulevard Maurice Bourdet 13001 Marseille, France;
loann.DESBOULETS@univ-amu.fr

November 3, 2020

Abstract

This paper is devoted to practical use of the Manifold Selection method presented in Desboulets (2020). In a first part, I present an application on financial data. The data I use are continuous futures contracts underlying commodities. These are multivariate time series, for the period 1985-2020. Representing correlations in financial data as graphs is a common task, useful in Finance for risk assessment. However, these graphs are often too complex, and involve many small connections. Therefore, the graphs can be simplified using variable selection, to remove these small correlations. Here, I use Manifold Selection to build sparse graphical models. Non-linear manifolds can represent interconnected markets where the major drivers of prices are unobserved. The results indicate the market is more strongly interconnected when using non-linear manifold selection than when using linear graphical models. I also propose a new method for filling missing values in time series data. I run a simulation and show that the method performs well in case of several consecutive missing values.

Keywords – Non-parametric, Non-linear Manifolds, Variable Selection, Neural Networks

1. Motivations

Linear model are often not flexible enough to depict correlations among random variables. The recent success of non-parametric approaches shows that more general model are better suited to represent the data. A simple statistical tool for representing these correlations is known as the "Graphical Model". However, a non-parametric approach requires a lot of data. And it is a common issue when collecting data, that part of them are missing. In this paper, I propose a new method for estimating non-parametric graphical models, in the context of financial markets. I solve the missing values problem by exploiting a non-parametric time series model, proposed by Zimmermann et al. (2005), to impute missing values. I compare this approach to others methods, and show that in simulations it has better performance in recovering the true missing values for time series data.

Graphical models are often used in Finance, from modelling interdependences of international markets (Wang, 2010; Lafferty et al., 2012; Arbia et al., 2018) to risk management (Millington and Niranjan, 2017). Taking into account contemporaneous correlations in the basis of optimal portfolio theory, and the graphical model is a simple visual tool in this respect. Here, I am interested in a specific type of assets called "Futures". Futures are standardized financial contracts meant to deliver a pre-specified amount of underlying, and is traded on the stock exchange for a future date. The delivery date can vary from a contract to another, and is referred to as its "maturity". The data I can observe on the market are the time series of prices and volumes for various maturities, where the underlying are various commodities. It is usual not to consider prices but rather their changes, called the returns, to avoid non-stationarity issues. It is very likely that the returns of a single commodity at different maturities are linearly correlated. There is no reason that the price of copper for a delivery in 2 months vary very differently from its price for a delivery in 3 months. However, the linearity assumption might not hold for correlations between different commodities, or between returns and volumes. Therefore, the question is: are there non-linear correlations between futures returns and volumes of different commodities? I will consider a non-parametric graphical model to answer this. Also, to find the most important ones, I will use a sparse graphical model. A sparse model gets rid of insignificant/small correlations. That is useful in presence of high dimensional data, which is the case here, because non-parametric inference suffers the curse of dimensionality. This problem is then limited with a sparse model, because it lowers the number of parameters to be estimated.

First of all, a graphical model is based on a graph, denoted $\mathcal{G}$, which is composed of edges and vertices. Each vertices correspond to a random variable, while the edges represent the dependences that hold in that distribution. Another information that is often added to the graph, is the strength of the dependences. That can be represented by the thickness of the edge. The thicker the edge, the stronger the correlation. One of the simplest linear graphical model is the Gaussian Model

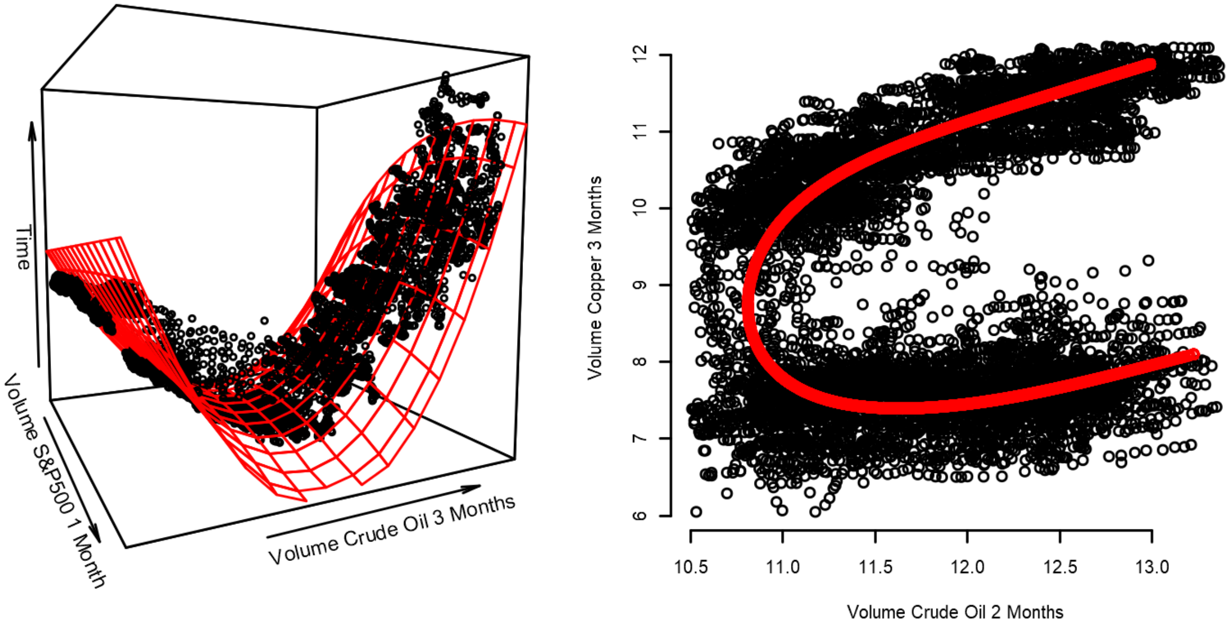
$$\mathbf{X} \in \mathbb{R}^p \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

The graph is given by the inverse linear covariance matrix $\boldsymbol{\Sigma}^{-1}$, also called the precision matrix. This matrix is used as the adjacency matrix of the graph, the vertices represent the variables, and an edge between the variables X_i and X_j is drawn if $\boldsymbol{\Sigma}^{-1}_{i,j}$ is non-zero. Using an empirical estimate of $\hat{\boldsymbol{\Sigma}}$ leads to fully connected graphs. Indeed, it is numerically improbable to obtain exact zeros in $\hat{\boldsymbol{\Sigma}}^{-1}$. That is why a sparse graphical model is required. Estimating a sparse graphical model turns out to be equivalent to solving a variable selection problem (Banerjee et al., 2008). One very well-known solution to variable selection has successfully been applied to sparse graphical modelling is the LASSO (Tibshirani, 1996). The resulting method is known as Graphical LASSO (Friedman et al., 2008). It operates variable selection on the linear precision matrix. So solving the Graphical LASSO accounts only for linear dependences. A non-parametric extension to the Graphical LASSO is the NonParanormal distribution (hereafter NPN) of Lafferty et al. (2012). The authors assume that a non-parametric, but monotonic, transformations $f_j(X_j)$ of the data follow the Gaussian Graphical Model.

$$\mathbf{f}(\mathbf{X}) = (f_1(X_1), f_2(X_2), \cdots, f_p(X_p)) \sim \mathcal{N}(\mu_{\mathbf{f}}, \boldsymbol{\Sigma}_{\mathbf{f}}). \tag{1}$$

The Graphical LASSO is then applied to $\boldsymbol{\Sigma}_{\mathbf{f}}$. The resulting graph is more general than the one from Graphical LASSO. Still, it is not fully non-parametric because it only allows monotonic transformations. It can exist complex dependences that will not be taken into accounts. In their paper, the authors propose a another fully non-parametric approach but the graph is not arbitrary sparse. It has to be acyclic, aka a tree. Despite being very useful in certain cases, in my case it is unlikely that the true correlations form a tree. So I only compare to their first approach, which is NPN. A more general graphical model has to allow for non-monotonic transformations. For instance, consider non-monotonic transformations of two random variables $(f_1(X_1), f_2(X_2))$. This means that their dependence is $X_2 = f_2^{-1} \circ f_1(X_1)$. The term $f_2^{-1} \circ f_1$ is called a function composition. And in the multivariate case, the distribution of the data having these kind of dependences is characterised by what is called

Fig. 1. Example of two non-linear manifolds estimated on real data



Notes: The manifolds are estimated via Non-Linear Least Squares (left) and Principal Curves (right). The data are Continuous Futures Contract downloaded from Quandl API, for the period 1985 - 2020, and further presented in section 3.1.

a "manifold". The equation of a manifold $\mathcal{M}$ is written as a composition of two functions f and g.

$$\begin{aligned} \mathcal{M} :=& \{\mathbf{x} \in \mathbb{R}^p | \mathbf{x} = f \circ g(\mathbf{x})\} \\ g : \mathbb{R}^p \rightarrow \mathbb{R}^r&, f : \mathbb{R}^r \rightarrow \mathbb{R}^p, p > r \end{aligned} \tag{2}$$

Also it does not impose conditional normality. Already from visual inspection, I have been able to observe non-linear manifolds in the data, as illustrated in Figure 1. The figure on the left is a striking example of non-linearity, while the picture on the right shows a dependence that is more complex than a function. These are contemporaneous data, i.e. at the same date. I do not expect to find such strong correlations with cross-correlations on the long run, otherwise there would be many possible arbitrages on the market. The manifold can also be interpreted using a latent structure τ, which represents its coordinates. The data are simply projections of these coordinates into the ambient $\mathbb{R}^p$ space.

$$\begin{aligned} & \mathbf{X} = f \circ g(\mathbf{X}) \\ \Leftrightarrow \quad & \mathbf{X} = \mathbf{f}(\tau) = \mathbb{E}[\mathbf{X} | \tau_{\mathbf{f}}(\mathbf{X}) = \tau] \end{aligned}$$

where τ is the low dimensional coordinate of greatest length, such that the reconstruction best fit the data:

$$\tau_{\mathbf{f}}(\mathbf{X}) = \sup_{\tau} \left\{ \tau : \|\mathbf{X} - \mathbf{f}(\tau)\|_F = \inf_{\mu} \|\mathbf{X} - \mathbf{f}(\mu)\|_F \right\}$$

And the latent structure represents unobserved variables, correlated to the observed ones, through non-linear functions. From an economic viewpoint, τ is proxy for unobserved variables such as decision-making processes of agents intervening in the market, regime changes, etc... The observed variables are only the prices and purchased volumes of futures contracts. It is very natural to assume that prices' movements and volumes are determined by common agents' decisions. If these decisions involve some non-linearities, then there will be non-linear manifolds in the data.

Then, there is the missing data issue. It is usual to encounter this problem when working with public sources, such as Quandl which is my source here. One solution could be to look for alternative sources with complete data. But the overall proportion of missing values in the downloaded data is $\frac{\text{total missing observations}}{\text{total observations}} = 4.2\%$. Since there is not that much missing values, I can go for another solution that consists in modelling the data, and impute the missing part based on that model. Because the data is a time series, I use a state-space model called Dynamically Consistent Neural Network (DCNN), proposed by Zimmermann et al. (2005). This model produces a complete time series that replicates exactly the prices and volumes at observed values. There also exist several other alternatives to model the data, reviews can be found in Harel and Zhou (2007) and Murray (2018). Contrary to the existing methods, the DCNN approach takes into account the dynamics of the series and is fully non-parametric. Still, I compare different alternatives on simulated time series and look for the one that best recover the true missing values. The results of these simulations indicates that DCNN performs better than any alternatives.

The rest of the paper is organised as follow. In section 2, I develop two graphical model estimators based on the two variable selection estimators proposed in Desboulets (2020). In section 3, I present the data and how to impute missing values. I use simulations to compare existing methods and the DCNN approach. In section 4, I present empirical estimates of the graphical models and compare the different estimators. As a result, non-linear graphical models show more inter-connected markets than linear graphical models. The manifold graphs show there is a strong dependence between returns and volumes of different commodities. This finding suggest that these dependences are very non-linear and not well represented as functions but rather as manifolds.

2. Sparse Graphical Models via Selection on Manifolds

In order to obtain a graph with non-linear manifolds, I use the Local Linear Manifold Selection (LLMS)[1] and the Diagonal Auto-Encoders Manifold Selection (DAMS) proposed in Desboulets (2020). I recall their definitions:

$$\text{LLMS} \Leftrightarrow \begin{cases} \mathbf{N}_{i,k}(\mathbf{x}_i) = \left\{ \mathbf{x} \in \mathbf{X} : \ \|\mathbf{x} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_i\|_{(k)} \right\} \\ \mathbf{R}_{\mathbf{N}_{i,k}} = Cor\left(\mathbf{N}_{i,k}\right) \\ \mathbf{R}_{\mathbf{N}_{i,k}} = (\mathbf{V}\mathbf{L}\mathbf{V}')_{\mathbf{N}_{i,k}} \\ \mathbf{V}_{\mathbf{N}_{i,k}} = \mathbf{V}_{\mathbf{N}_{i,k}} \mathbb{1}_{|(\mathbf{V}\mathbf{L}^{1/2})_{\mathbf{N}_{i,k}}|>\theta} \\ \hat{\mathcal{S}}_{\mathbf{N}_{i,k},\theta} = \left\{ j | \mathbf{V}_{\mathbf{N}_{i,k}}(j,s) \neq 0, \forall s \neq j \right\} \\ \Psi_{\mathbf{N}_{i,k},\theta} = \mathbb{1}_{j \in \hat{\mathcal{S}}_{\mathbf{N}_{i,k},\theta}} \\ \mathcal{P}_\theta = \frac{1}{n} \sum_{i=1}^{n} \Psi_{\mathbf{N}_{i,k},\theta} \\ \theta^* = \underset{\theta}{argmax} \left(Var\left[\mathcal{P}_\theta\right] \right). \end{cases}$$

$$\text{DAMS} \Leftrightarrow \begin{cases} \tilde{\mathbf{X}} = \mathbf{X}\mathbf{W_s} \\ \mathbf{Z^{(0)}} \equiv \tilde{\mathbf{X}} \\ \mathbf{Z^{(l)}} = tanh(\mathbf{Z^{(l-1)}}\mathbf{W_l}), \quad \forall l \in 1, \cdots, n_h - 1 \\ \hat{\mathbf{X}} = \mathbf{Z^{(n_h-1)}}\mathbf{W_{n_h}} \\ \tilde{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{W_s} \\ \mathbf{E} = \mathbf{X} - \tilde{\mathbf{X}} \\ \mathbf{W_s} = \underset{\mathbf{W_1},\ldots,\mathbf{W_{n_h}},\mathbf{W_s}}{argmin} \ Trace(\mathbf{E}'\mathbf{E}). \end{cases}$$

As shown by (Banerjee et al., 2008), a sparse graph can be obtained via variable selection. However passing from one to the other depends on the selection operator. A graph is represented by a $p \times p$ square matrix (the so-called adjacency matrix), but both LLMS ans DAMS outputs selection probability vectors of length p. The LLMS estimator gives $\mathcal{P}_{\theta^*}$ and the DAMS estimator gives $diag(\mathbf{W_s})$. So I have to modify the output in order to obtain a $p \times p$ square matrix. Also, because the graph is undirected, that square matrix has to be symmetric. For LLMS, the modification is simple. In the original algorithm, a sparse

[1] The second version as defined in Algorithm **??**, with Random-kNN.

correlation matrix is estimated for each neighbourhood. These sparse matrices are reduced to vectors and turned into the boolean form. Here, I just simply skip the reduction part, and directly turn the sparse matrices into the boolean form. So Ψ is no longer a vector, but a square matrix, that is used as the adjacency matrix. Then, I keep the original algorithm and average the results over all neighbourhoods. The LLMS estimator for the graph is therefore:

$$\hat{\mathcal{G}}_{LLMS} \to \sum_{i=1}^{n} \Psi_{\mathbf{N}_{i,k},\theta}.$$

Turning the DAMS estimator into a graph is more complicated. The selection weight is a diagonal matrix, and so it does not bear any information about pairwise dependences. A good candidate to get that information, is the symmetric square matrix of all partial derivatives $\frac{\partial X}{\partial \hat{X}'}$. These are obtained by the chain rule, as a product of the derivatives of each layers. Note that, it is a non-parametric analogue to the precision matrix Σ^{-1}. Indeed, it measures the strength of the pairwise conditional dependences. But the derivatives are not constants, they are vectors. To summarize them, the simplest is to average. However, it might be that some dependences are non-monotonic. Therefore, the derivative of a non-monotonic function has positive and negative values. And these values will cancel out if averaged. To avoid that problem, I average the absolute values of the derivatives. In the end, I have a symmetric square matrix that measures pairwise non-parametric correlations. Still, it does not provide selection. The latter is brought by the vector $diag(\mathbf{W_s})$. To combine both, I would like to multiply them element by element. But their dimensions are not compatible. A simple way to achieve this is turn the selection vector into a matrix as well: $\tilde{\mathbf{W}} = (\mathbf{diag}(\mathbf{W_s})\mathbf{diag}(\mathbf{W_s})')^{1/2}$. $\tilde{\mathbf{W}}$ takes on continuous values between 0 and 1. In the original version of DAMS, there is only one output. There is no penalty introduced, that is handled by the network itself. This means that here, it will only produce one graph. To get fair comparison with the other methods, I introduce a threshold to obtain a sequence of graphs. Using a threshold is equivalent to say: "an edge between two variables is drawn if and only if the two selection weights associated are greater than some threshold". The DAMS estimator for the graph is therefore:

$$\hat{\mathcal{G}}_{DAMS} \to \tilde{\mathbf{W}} \mathbb{1}_{\tilde{\mathbf{W}}>\theta} \sum_{i=1}^{n} \left| \prod_{j=n_h}^{1} (1 - tanh(Z^{(j)})^2)' W_j \right|$$

All methods then give sequences of graphs. These sequences vary with a penalty parameter. Up to now, there is little work on an optimal choice for the penalty. Theoretical results were provided by Knight and Fu (2000); Zhao and Yu (2006); Bickel et al. (2009). However the problem of estimating it is always ill-posed. Practical solutions were given by Belloni et al. (2011) and Chichignoud et al. (2016), but both are quite computationally intensive. Yet, I gave an optimal penalty choice for LLMS, which has shown to give good results in simulations. Also, to make fair comparison, it is usual in the literature to choose the graphs so that they have the same number of vertices. What I will do then, is to chose the penalty parameter of other methods, such that their graph has the same sparsity as LLMS. I will also comment the full sequences of graphs.

3. Empirical Data

3.1. Presentation

I focus on commodity market data, and more precisely on Futures Contracts (FC) of the oil and metal markets. These two markets are interesting because they are heavily traded, and directly connected to the real economy. The underlying of each contract is a commodity, used in the industry for a wide variety of goods. Therefore, rational decisions taken by economic agents, can induce patterns in the prices and volumes of FC. Keep in mind that I only look at contemporaneous correlations, there is no predictive power nor any possible arbitrage from these correlations. Already in Figure 1, I can see there are strong non-linear correlations. Nonetheless, having raw futures contracts is not suitable for long term analysis, since they are traded for a short amount of time. For instance, the contract ”Copper April 2020” starts to be traded usually 6 months in advance i.e. November 2019, and wont be traded after April 2020 as it ceases to exist. The solution is to use continuous FC. These are chains of individual FC. The FC for a given commodity are combined into long term time series, each with a different maturity, expressed in months. I collect these data from Quandl API, for the period 1985 - 2020. There are different kind of assets: major commodities from the Metal Market, major commodities of the Oil Market. On top of that I also include the S&P500 which serves as market index, and the US Treasury Bond (USTBill) as an interest rate index. In total, I have 7618 daily prices and volumes, for different maturities. For some commodities I have up to 6 months ahead contracts. In the end, I have 58 series for constructing a graph.

The prices are not used as is, I prefer to use returns as it is usually done in others studies. The volumes are also expressed in log, to reduce their range. The series I obtain are very homogeneous in terms of moments, as described in Tables 1 and 2. Each series is described by its name, an underscore, and the maturity of the contract in months. They mostly have the same moments, especially the different maturities of a given commodity. From visual inspection, prices follow very similar paths, which explains the similar moments. That supports the idea that the returns at different maturities of a single commodity are greatly and linearly correlated.

Table 1: Descriptive Statistics - Continuous Futures Contracts (returns)

Returns	Mean	Sd	Min	Q1	Median	Q3	Max	Skewness	Kurtosis
COPPER_1	0.00	0.02	-0.15	-0.01	0.00	0.01	0.16	-0.16	11.18
COPPER_2	0.00	0.02	-0.16	-0.01	0.00	0.01	0.20	0.21	13.78
COPPER_3	0.00	0.02	-0.17	-0.01	0.00	0.01	0.18	0.11	13.41
COPPER_4	0.00	0.02	-0.15	-0.01	0.00	0.01	0.24	0.31	16.72
GOLD_1	0.00	0.01	-0.16	-0.00	0.00	0.01	0.16	0.14	28.84
GOLD_2	0.00	0.01	-0.09	-0.00	0.00	0.01	0.10	0.06	12.04
GOLD_3	0.00	0.01	-0.09	-0.00	0.00	0.01	0.10	-0.05	11.64
GOLD_4	0.00	0.01	-0.16	-0.00	0.00	0.01	0.16	0.13	26.71
GOLD_5	0.00	0.01	-0.11	-0.00	0.00	0.01	0.10	-0.03	11.35
GOLD_6	0.00	0.01	-0.09	-0.00	0.00	0.01	0.08	-0.06	9.99
SP500_1	0.00	0.01	-0.11	-0.00	0.00	0.01	0.11	-0.42	11.66
SP500_2	0.00	0.02	-0.15	-0.00	0.00	0.01	0.16	0.08	21.83
SP500_3	0.00	0.01	-0.18	-0.00	0.00	0.01	0.19	0.05	27.51
USTBILL_1	0.00	0.01	-0.20	-0.00	0.00	0.00	0.10	-2.45	88.88
USTBILL_2	0.00	0.01	-0.19	-0.00	0.00	0.00	0.10	-2.14	70.70
SILVER_1	0.00	0.02	-0.20	-0.01	0.00	0.01	0.18	-0.45	11.10
SILVER_2	0.00	0.02	-0.20	-0.01	0.00	0.01	0.13	-0.50	8.70
SILVER_3	0.00	0.02	-0.19	-0.01	0.00	0.01	0.18	-0.41	12.39
PALLADIUM_1	0.00	0.03	-0.34	-0.01	0.00	0.01	0.37	0.02	27.59
PLATINUM_1	0.00	0.02	-0.33	-0.01	0.00	0.01	0.22	-0.96	38.01
CRUDEOIL_1	0.00	0.02	-0.21	-0.01	0.00	0.01	0.26	0.11	11.01
CRUDEOIL_2	0.00	0.02	-0.12	-0.01	0.00	0.01	0.15	-0.15	6.37
CRUDEOIL_3	0.00	0.02	-0.12	-0.01	0.00	0.01	0.13	-0.16	6.30
GASOIL_1	0.00	0.02	-0.17	-0.01	0.00	0.01	0.18	-0.05	9.13
GASOIL_2	0.00	0.02	-0.20	-0.01	0.00	0.01	0.21	-0.00	9.40
GASOIL_3	0.00	0.02	-0.12	-0.01	0.00	0.01	0.11	0.04	6.15
GASOIL_4	0.00	0.02	-0.11	-0.01	0.00	0.01	0.10	-0.02	6.17
GASOIL_5	0.00	0.02	-0.10	-0.01	0.00	0.01	0.10	-0.02	5.53
GASOIL_6	0.00	0.02	-0.11	-0.01	0.00	0.01	0.10	-0.02	6.48

Table 2: Descriptive Statistics - Continuous Futures Contracts (log-volumes)

log(Volumes)	Mean	Sd	Min	Q1	Median	Q3	Max	Skewness	Kurtosis
COPPER_1	7.34	1.11	0.69	6.60	7.32	8.03	10.55	0.15	3.37
COPPER_2	8.98	1.43	1.10	7.81	8.38	10.32	12.11	0.31	2.09
COPPER_3	9.08	1.74	0.69	7.55	8.24	10.77	12.10	0.21	1.58
COPPER_4	8.74	1.75	0.61	7.22	8.07	10.39	12.32	0.25	1.78
GOLD_1	8.76	2.77	0.69	6.45	8.79	11.34	13.22	-0.23	1.88
GOLD_2	10.99	1.47	1.95	10.12	11.28	12.21	13.19	-1.19	5.22
GOLD_3	10.16	1.27	1.10	9.38	10.07	10.82	13.19	-0.31	5.93
GOLD_4	9.61	0.87	5.86	8.99	9.62	10.17	12.70	0.12	2.97
GOLD_5	9.17	0.76	5.39	8.64	9.20	9.70	11.29	-0.17	3.10
GOLD_6	8.81	0.78	5.94	8.26	8.82	9.41	10.98	-0.06	2.91
SP500_1	12.29	0.81	8.46	11.83	12.24	12.97	13.41	-0.63	2.88
SP500_2	9.12	2.03	0.69	8.34	9.20	10.25	13.41	-1.00	5.47
SP500_3	7.20	1.62	0.69	6.66	7.67	8.23	10.05	-1.52	5.64
USTBILL_1	12.49	1.43	0.69	12.25	13.00	13.38	13.96	-2.08	8.67
USTBILL_2	10.19	2.66	0.69	8.98	10.40	12.52	13.84	-0.88	3.47
SILVER_1	7.25	3.08	0.69	5.06	6.91	10.51	12.06	-0.17	2.06
SILVER_2	7.89	3.40	0.69	5.07	9.32	10.80	12.14	-0.63	2.05
SILVER_3	9.33	2.53	0.69	9.04	10.26	10.90	12.09	-1.84	5.63
PALLADIUM_1	7.01	2.53	0.69	5.31	7.72	8.96	10.65	-0.79	2.71
PLATINUM_1	7.58	2.70	0.69	5.31	8.77	9.49	11.35	-0.74	2.45
CRUDEOIL_1	11.80	0.85	6.01	11.26	11.78	12.54	13.40	-0.44	3.02
CRUDEOIL_2	11.82	0.67	10.30	11.24	11.82	12.39	13.36	0.07	1.94
CRUDEOIL_3	11.21	0.65	9.77	10.65	11.12	11.75	12.80	0.33	2.03
GASOIL_1	10.56	0.98	2.56	10.08	10.55	11.26	12.56	-1.39	8.01
GASOIL_2	10.66	0.90	0.69	10.02	10.53	11.43	12.56	-0.31	6.26
GASOIL_3	10.18	0.91	6.26	9.47	10.05	10.96	12.26	0.15	2.15
GASOIL_4	9.68	0.92	6.26	8.98	9.64	10.40	12.06	0.09	2.50
GASOIL_5	9.36	1.01	5.43	8.61	9.35	10.12	11.90	-0.01	2.59
GASOIL_6	9.12	1.09	5.11	8.32	9.17	9.94	11.85	-0.11	2.79

3.2. Fill missing values in the data

The overall proportion of missing values in the downloaded data is 4.2%. But the proportion of *pairwise* missing values is $\frac{\text{number of days with at least one missing observations}}{\text{total number of days}} = 46.4\%$. That is an issue for LLMS for 2 reasons: (1) local correlation matrices are biased when there is a large proportion of missing values in the neighbourhood (2) it biases the computation of distances. The pairwise missing value problem is much less an issue for DAMS, as it is a global estimator. Still, it is better to have more observations. So, there are incentives to complete the data.

It is usual to assume the data are "Missing At Random" (MAR). This is convenient to assume, because that means one does not have to model the process that causes the missing values. Also it means one can estimate an unbiased, but less precise, model on the incomplete data, and then use it to impute the missing values. In my case, MAR is a reasonable assumption for two reasons. First, the missing parts of the data are probably due to errors in Quandl itself when it builds the chains of FC. There is no reason to believe this occurs for specific time or asset, it should be random. Second, if you remove 5% of the data at random, you obtain in average 40% of pairwise missing values, this is consistent with the percentage I have. So I can fit a model on incomplete data to impute the missing values.

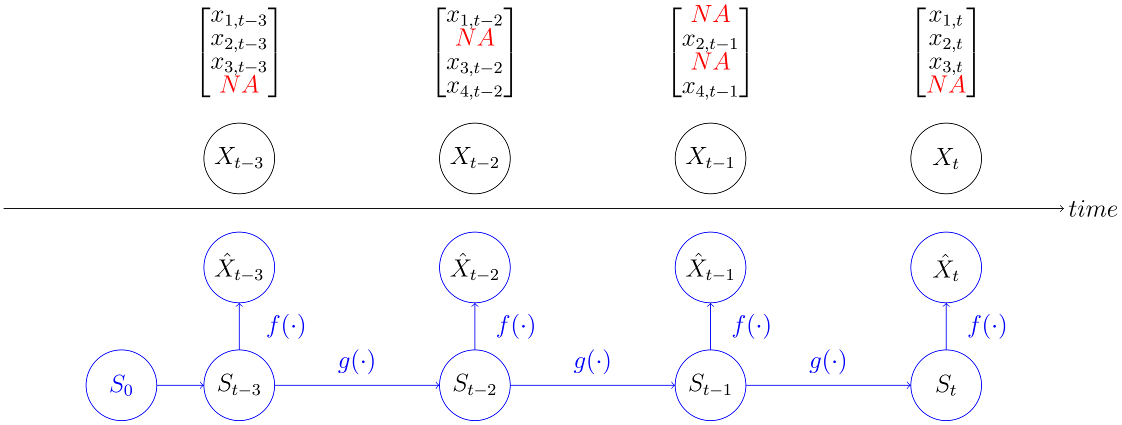
3.2.1. The Dynamically Consistent Neural Network approach

The Dynamically Consistent Neural Network of Zimmermann et al. (2005) is a state-space model, and was originally designed for forecasting. The main assumption it makes is there exists a deterministic process that has produced the entire series, with no external influences, aka a closed dynamical system. That is very useful for predictions, especially at long-term. The closed system is composed of two different components, the first being the observed data X_t, and the second the unobserved data S_t. These unobserved vectors play two roles: (1) they constitute the hidden layers of the network that approximates the dynamics, and (2) they form hidden variables that close the system perfectly. There is no a priori on the form of the time dynamics, the model is fully non-parametric. The DCNN is represented as state-space model, with unknown transition functions $f(\cdot)$ and $g(\cdot)$:

$$\begin{aligned} S_t &= g(S_{t-1}), \\ X_t &= f(S_t). \end{aligned} \tag{3}$$

Here, I use DCNN as a generating process to construct an exact copy of the observed data. And since the model is based only on an initial state, the reconstructed series are complete series. The missing value problem is therefore solved. However there exist multiple solutions that solve the DCNN. To take this problem into account, I run the DCNN several times and obtain a distribution of possible values. The imputed value in the end is the median of that distribution. The DCNN imputation method is illustrated in Figure 3.2.1.

Fig. 2. Missing values imputation via DCNN

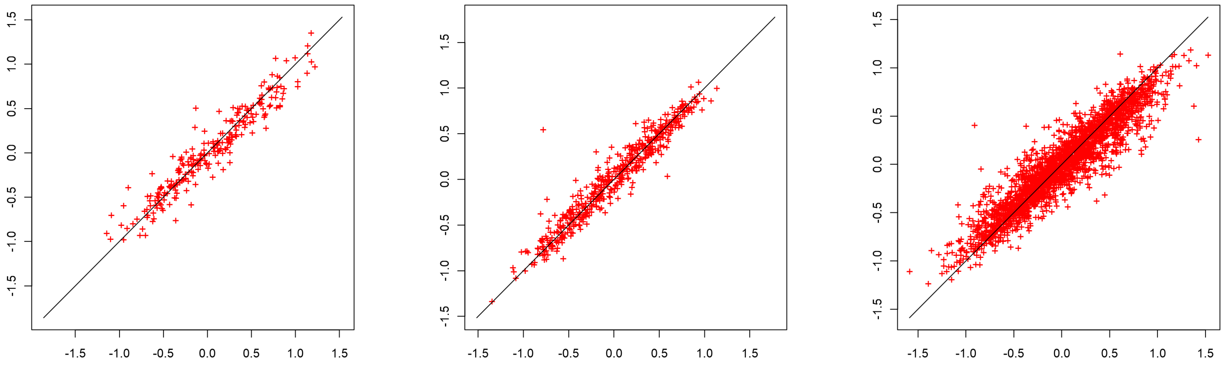


3.2.2. Performance of filling methods through simulations

I run a simulation to test if DCNN is a valid method for imputation. I will also compare it with more standard methods, such as median-value imputation and previous-value imputation. These are unlikely to achieve what I want, which is to recover the ”true” missing values. Previous-value imputation implies static prices from one day to the other. But for some series, there are several missing values in a row. It is unlikely that prices stagnate for several days. I will also compare with 3 more recent methods: Predictive Mean Matching (PMM) of Vink et al. (2014), MissForest of Stekhoven and Bühlmann (2012) and k-Nearest Neighbours of Troyanskaya et al. (2001). They are all available on the software R, respectively from packages `mice`, `missForest` and `impute`. These are more complex than median-value and previous-value, they make use of multivariate information to interpolate the data.

I simulate a multivariate time series X_t and S_t according to equation 3, with $\begin{pmatrix} f(x) \\ g(x) \end{pmatrix} = \mathbf{A}tanh(\mathbf{B}x)$ where $\mathbf{A}$ and $\mathbf{B}$ are random square matrices uniformly distributed on $[-5, +5]$. This parametrization is convenient as it can produce any smooth non-linear functions f and g, but always different at each run. I only keep the series X_t, and I remove a certain percentage of the data. These series are standardized to have mean zero and standard deviation one. I chose two sample sizes, 500 and 5000, and $p = 50$ variables. The ensemble is made of 100 networks. The different levels of overall missing values are 5%, 10% and 50%. This corresponds, in average, to pairwise missing values levels of 40%, 65% and 100% respectively. That is in line with our data, that has 4.2% overall and 46.4% pairwise missing values. Thus, it supports the Missing At Random hypothesis. The results of the simulations are reported in Figures 3 and 4, and in Table 3.

Fig. 3. Comparison of estimated versus true missing values, $n = 500$

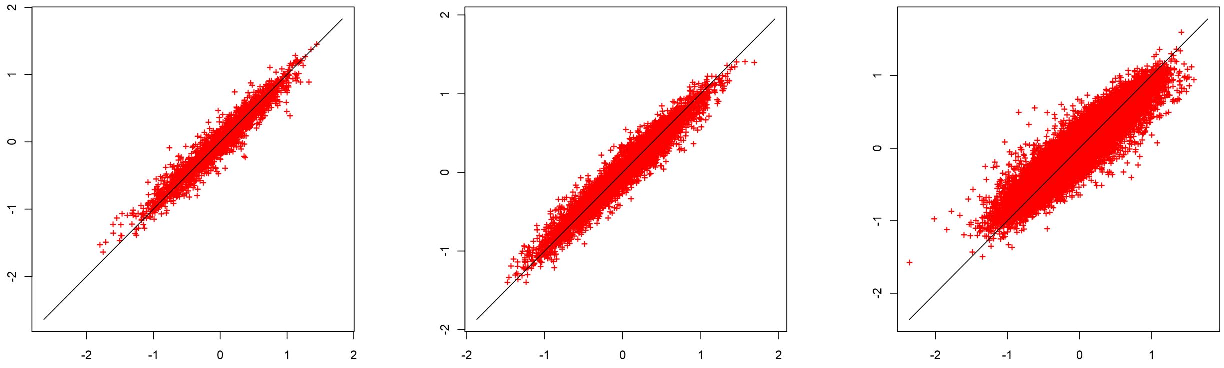


(a) 5%

(b) 10%

(c) 50%

Fig. 4. Comparison of estimated versus true missing values, $n = 5000$



(a) 5%

(b) 10%

(c) 50%

The plots indicates that DCNN is a unbiased estimator of the missing values. The recovered values lies around the 45^o line. The sample size does not seem to have a great effect on the efficiency of the procedure. However the performance differ a bit w.r.t the percentage of missing values. The more values are missing the less precise DCNN is. But when comparing mean squared errors in the table, DCNN outperforms all other methods. The benchmark for comparison should certainly be the median imputation, as most series are symmetric. On average its MSE is around 0.25. The previous value imputation is certainly the worst, having a MSE over 0.45. Still, the three alternatives are all better than the benchmark. PMM performs significantly better with more observations, and even with a high number of missing values. MissForest has also a great performance, but when the percentage of missing values is 5% or 10%. It is much less reliable at 50%. The last method, kNN, performs very well. This method seems to be only affected by the percentage of missing values, its performance is independent of sample size.

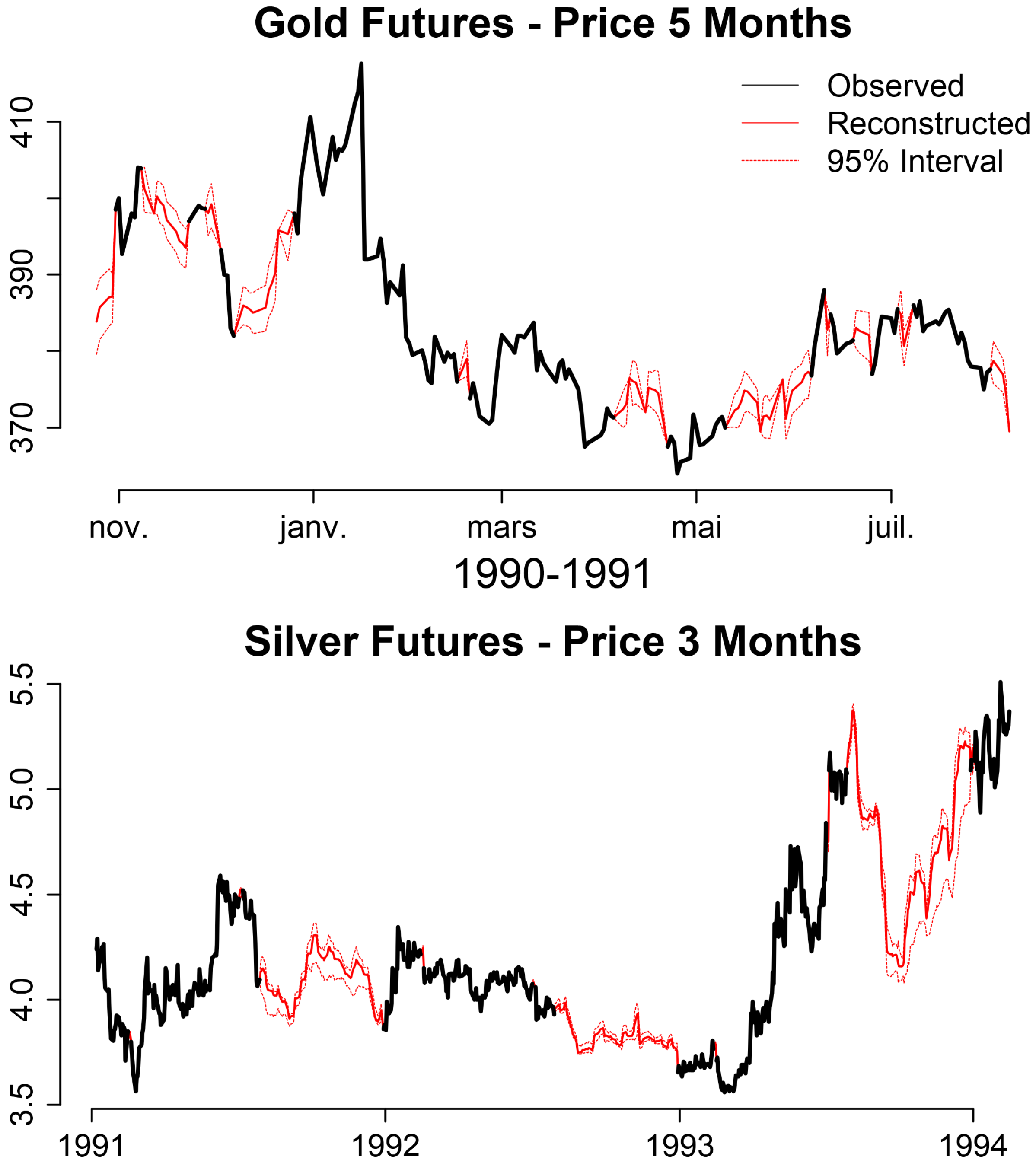
Table 3: Mean Squared Error of reconstructed series

	DCNN	Median	Previous	PMM	MissForest	kNN
5% missing						
$n = 500$	0.023	0.234	0.429	0.301	0.118	0.085
$n = 5000$	0.015	0.245	0.491	0.154	0.107	0.081
10% missing						
$n = 500$	0.021	0.246	0.464	0.196	0.129	0.144
$n = 5000$	0.015	0.251	0.454	0.185	0.113	0.135
50% missing						
$n = 500$	0.033	0.246	0.453	0.294	0.239	0.232
$n = 5000$	0.029	0.252	0.449	0.163	0.214	0.237

Overall, from these simulations, I prefer to use DCNN for filling missing values on my actual data. Figure 5 shows examples of two reconstructed series. These two series are the ones having the most missing values of all the dataset. The paths exhibit various patterns, visually very consistent with the other data. One should also note that the distribution is very narrow, the Ensemble of DCNN produces very similar paths.

If ever the imputation is in fact inconsistent with the true data, the graph made with original data may differ a lot from the graph made on reconstructed data. Because if the imputation is biased, the imputed data may not lie on a manifold. Keep in mind that DCNN never imposes a manifold constraint, data can have correlations, or not. They simply follow the dynamics. So, a biased imputation may affect the selection. On the opposite, if the reconstructed data is close to the true values, and that a manifold exist, the two graphs should not differ that much. So I will do both graphs, and compare them.

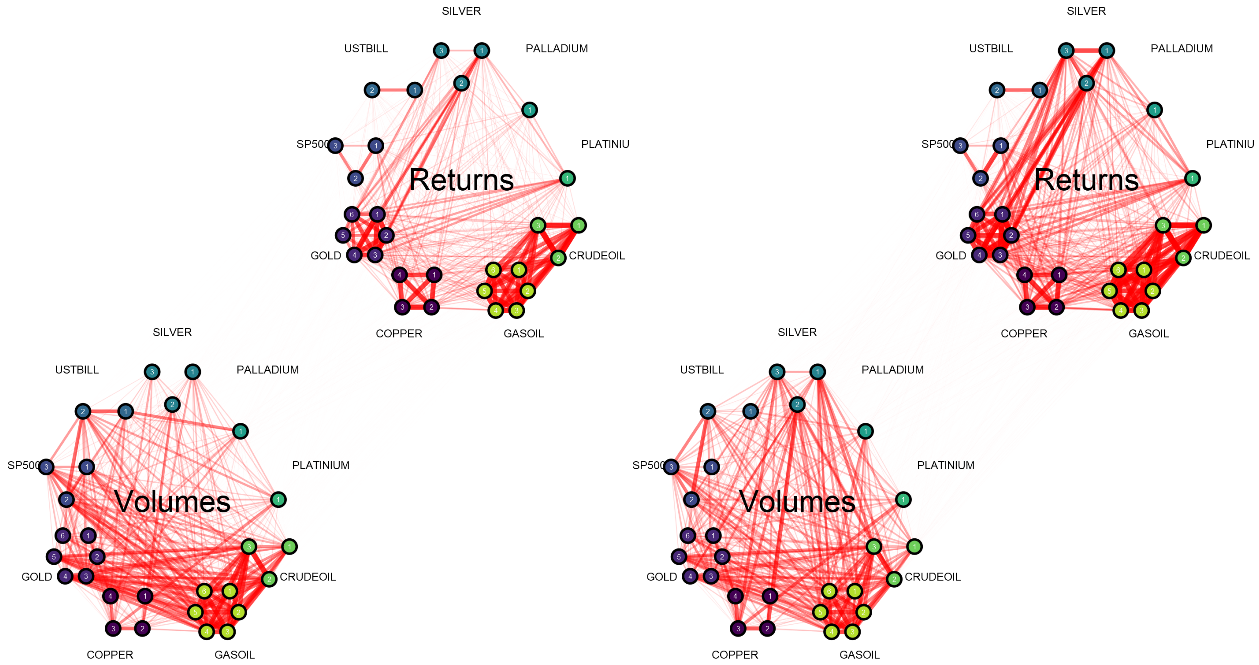
Fig. 5. Examples of reconstructed series via DCNN



4. Empirical estimates of the Sparse Graphical Models

Figure 6 shows the simple linear gaussian graph estimated with empirical covariance matrix. We observe great linear dependences among returns and among volumes, but nearly zero not in between.

Fig. 6. Gaussian Graphical Model



(a) Reconstructed Series

(b) Original Series

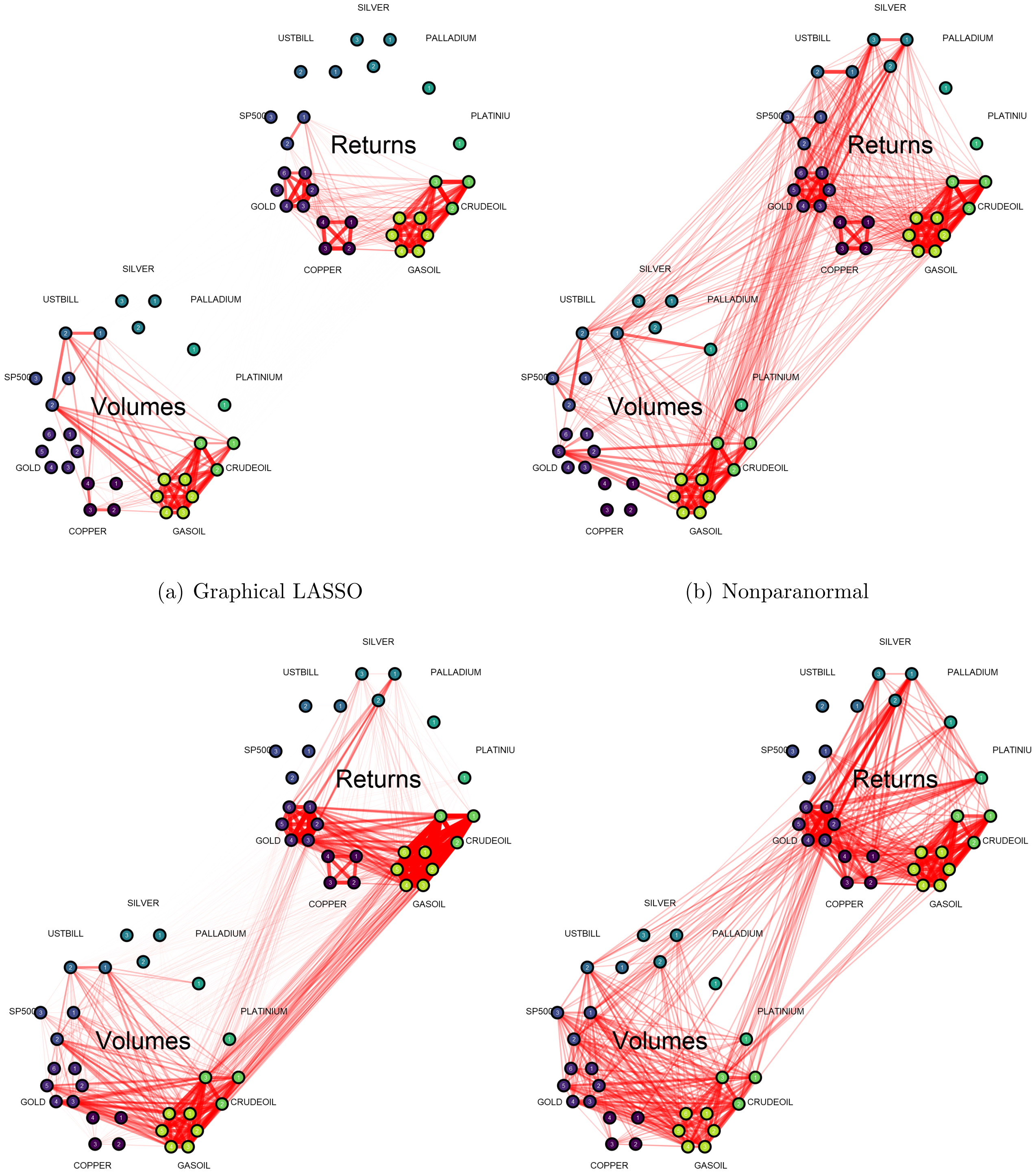
On the returns' side, the oil market (gas oil + crude oil) is strongly correlated for all maturities. Also copper exhibits strong correlations among all its different maturities. Gold shows the same dependences as copper, in addition with some stronger correlations to the 1 month and the 2 months silver and the 1 month platinum. All remaining dependences between returns all smaller (something) and concerns all maturities of gold, copper, gas oil, crude oil and silver.

On the volumes' side, there are overall as much correlations as in returns. However, for a given commodity, the dependences among its different maturities are very low. There are stronger links between commodities, compared to the returns. The interdependences between the oil market and 3 months up to 5 months gold quite high (something). I also observe strong dependences between the oil market and the 2 and 3 months S&P500.

The overall pattern shows dependences in majority for gold, copper, gas oil and crude oil. For returns these commodities are ”self-correlated”, there are strong links between the different maturities of a commodity but not much with other commodities. For volumes it is the opposite. The commodities are linked between each others but not much with themselves.Now, keep in mind that we only look at simple linear correlations. There might be dependences involving discontinuities, non-linearities, or maybe outliers.

The objective of a Sparse Graphical Model (SGM) is to reduce the complexity of the graph, by pushing down to zero the less significant dependences.
The results of all 4 estimated graphs are displayed in Figure 7. There is no simple optimal choice for the penalty parameter in Graphical LASSO and NPN. Therefore, there is an infinity of produced graphs for these two methods. However, there is a single graph produced by LLMS. Indeed, an optimal rule was proposed in Desboulets (2020) for the choice of the penalty and therefore for the choice of the optimal graph. To make fair comparison, as it is usually done in the literature, I choose the graphs for DAMS, Graphical LASSO and NPN, such that they have the same number of vertices. Based on this, I have chosen the penalty parameter so that all graphs have around 650 edges.

Fig. 7. Comparison of estimated graphs with ∼ 650 edges



(a) Graphical LASSO

(b) Nonparanormal

(c) Local Linear Manifold Selection

(d) Deep Autoencoders Manifold Selection

The graph of Graphical LASSO is shown in Figure 7(a). As I could have guessed, it resembles the Gaussian graphical model, having removed all low correlations. Only remains self-dependences for returns and dependences between the oil market and the 2 months S&P500 for volumes. There is also still a significant correlation between the volumes of the 2 months S&P500, mid-long term gold (4 to 6 months) and the US T-bill.

The graph of NPN is shown in Figure 7(b). It essentially has the same strong dependences as Graphical LASSO. Still, there are three differences. First, a lot of small dependences in the returns between, gas oil, crude oil, copper, gold, S&P500 and US T-bill. Second, there is no more connections between the oil market and the S&P500 for volumes. It might be due to the fact that these connections are lower than the ones that appeared on the returns' side. Because there is almost the same number of vertices, some connections from Graphical LASSO have been removed. Third, there are small correlations between returns and volumes. Indeed, the volumes of the oil market are connected to the returns of the oil market and short-term gold (1 to 3 months). There are also smaller dependences between the volumes of US T-bill and gold, copper and S&P500. This shows that taking into account non-linearities do modify the estimated graph.

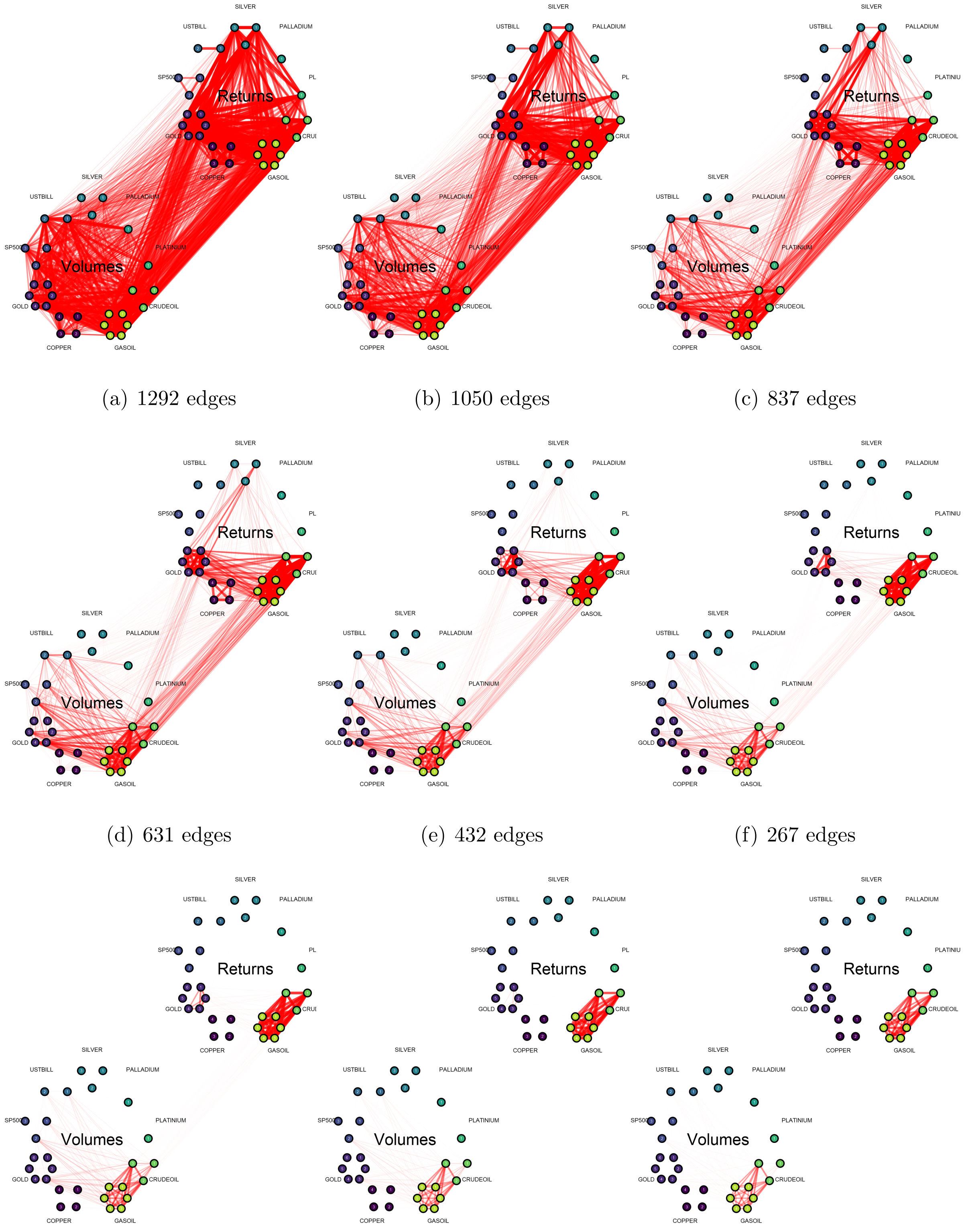
The graph of LLMS is shown in Figure 7(c). At first sight, it seems closer to the graph produced by Graphical LASSO than it is to NPN. Self-correlations among major commodities (gas oil, crude oil, copper and gold) are stronger than Graphical LASSO and NPN. The two main differences w.r.t to the two previous graphs are (1) stronger correlation between returns and volumes of the oil market and (2) a lot of moderate dependences among volumes of gas oil, crude oil, mid-term gold (3 to 5 months), 1 and 2 months S&P500 and US T-bill. The connections between gold returns and oil market volumes are weaker that it was for NPN, but still non-zero.

The graph of DAMS is shown in Figure 7(d). Contrary to LLMS, this one is closer to NPN than it is to Graphical LASSO. However it has more connections between volumes of the oil market and S&P500 for all maturities. The main difference w.r.t LLMS is the dependences of silver, platinum and gold returns, which appear to be very strong on this graph.

Moving on the full sequences of graphs. In relative terms, the comparison between all four methods would remain the same. When looking at intermediate levels of sparsity, for example 500 or 1000 edges, two method have nearly the same differences. The comments made before would not change that much. For low sparsity, around 200 edges, the graphs all become nearly the same: they exhibit strong correlations for gold, copper, crude oil and gas oil. These correlations are mainly between their own maturities. Only crude oil and gas oil have cross-correlations. The only noticeable difference is w.r.t Graphical LASSO, that never drawn an edge between returns and volumes. Whereas the three other method indicate intermediate to high correlations. This, is a strong indicator that there are non-linear correlations between returns and volumes of many commodities.

The graphs computed on the reconstructed data are also very similar with the graphs computed on the original data, that has missing values. The latter are available in Appendix A. The general appearance of the graphs, for all methods, remains the same. This result is an evidence that the imputation method I used is valid in my setting. However, for all methods, the graphs from original data exhibit slightly stronger correlations. The edges are slightly thicker. Two possible reasons for this. First, the missing value imputation has introduced some noise, and therefore reduced the strength of the correlations coefficients. Second, having less data points because of missing values has lead to over-fitting, and so the correlations are biased upward. Whatever the reason, the conclusion is that missing values imputation via DCNN seems to be reliable for time series.

Fig. 8. LLMS estimates for a range of penalty



(a) 1292 edges

(b) 1050 edges

(c) 837 edges

(d) 631 edges

(e) 432 edges

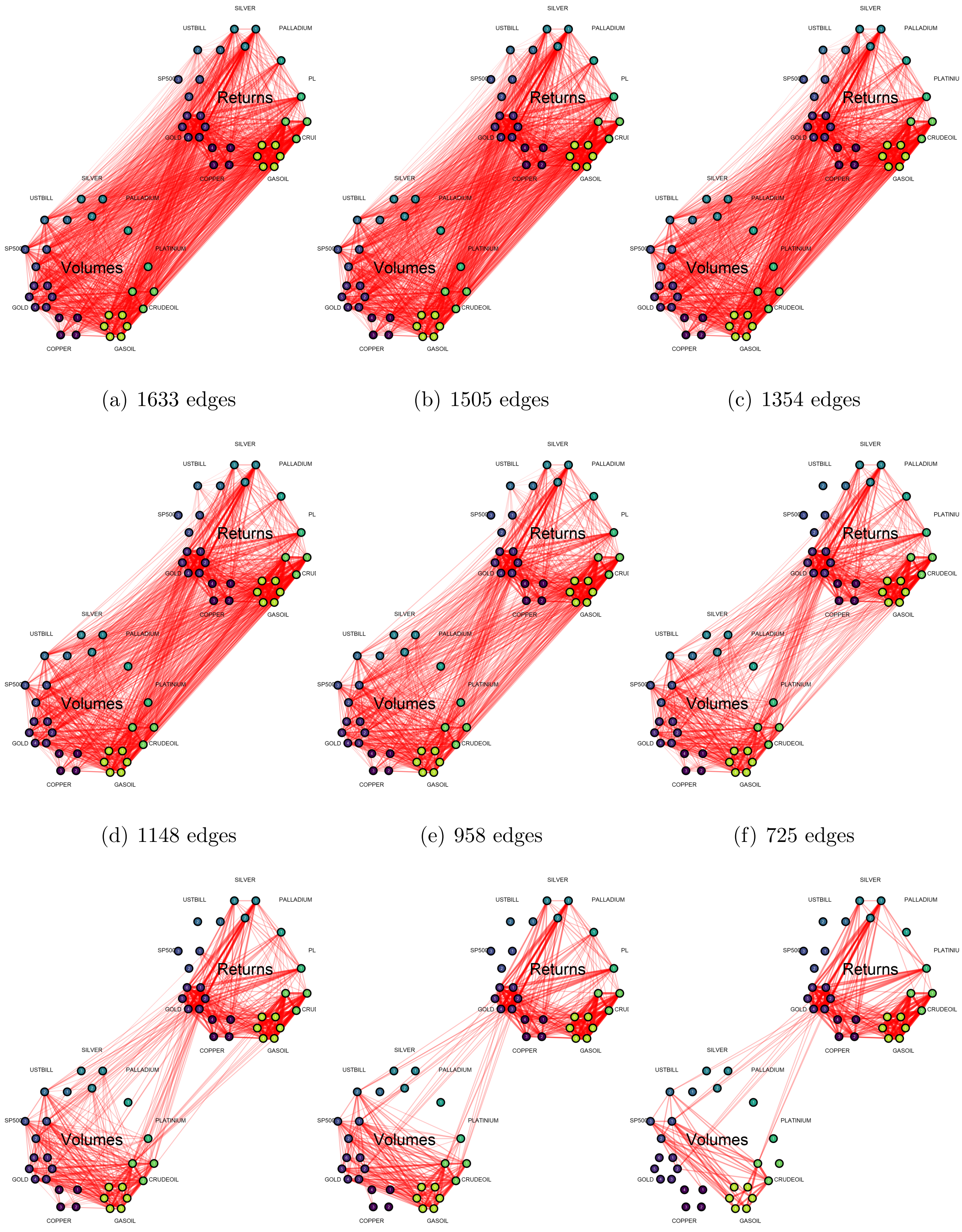
(f) 267 edges

(g) 170 edges

(h) 124 edges

(i) 91 edges

Fig. 9. DAMS estimates for a range of penalty



(a) 1633 edges

(b) 1505 edges

(c) 1354 edges

(d) 1148 edges

(e) 958 edges

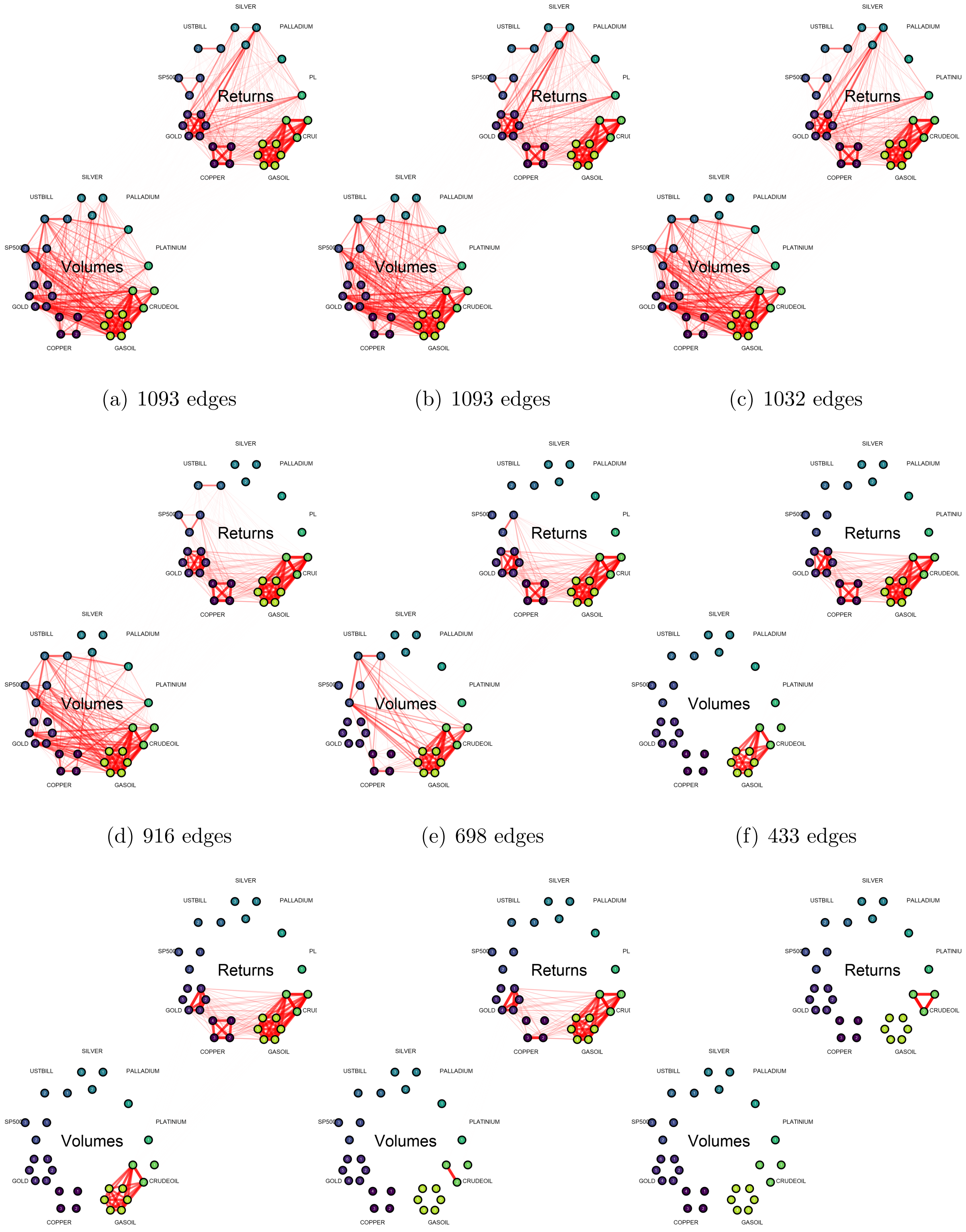
(f) 725 edges

(g) 545 edges

(h) 426 edges

(i) 257 edges

Fig. 10. Graphical LASSO estimates for a range of penalty



(a) 1093 edges

(b) 1093 edges

(c) 1032 edges

(d) 916 edges

(e) 698 edges

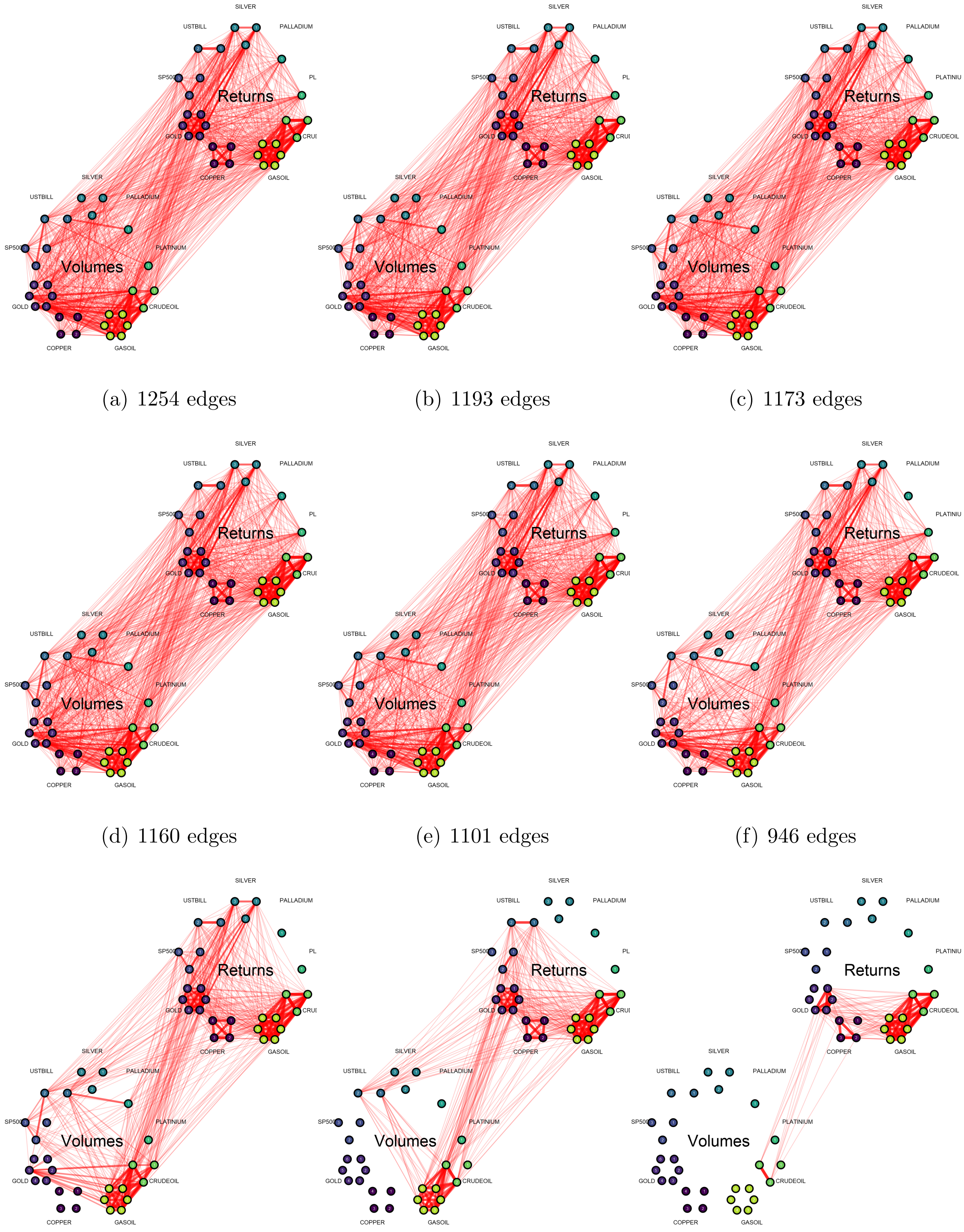
(f) 433 edges

(g) 291 edges

(h) 136 edges

(i) 6 edges

Fig. 11. Nonparanormal estimates for a range of penalty



(a) 1254 edges

(b) 1193 edges

(c) 1173 edges

(d) 1160 edges

(e) 1101 edges

(f) 946 edges

(g) 670 edges

(h) 501 edges

(i) 223 edges

5. Discussion

I presented here an application of "manifold variable selection" on financial data, using graphical models. The non-linear manifold graph is useful to represent more general correlations among time series. The results reveal more interconnected financial markets. They especially show many correlations between returns and volumes, that were non-existent in the other types of graphs. The difference among the two manifold selection estimators proposed, essentially lies in which assets are selected. The LLMS estimator highlight correlations on the oil market, while DAMS focuses on correlations between many assets' volumes and returns of the metal market.

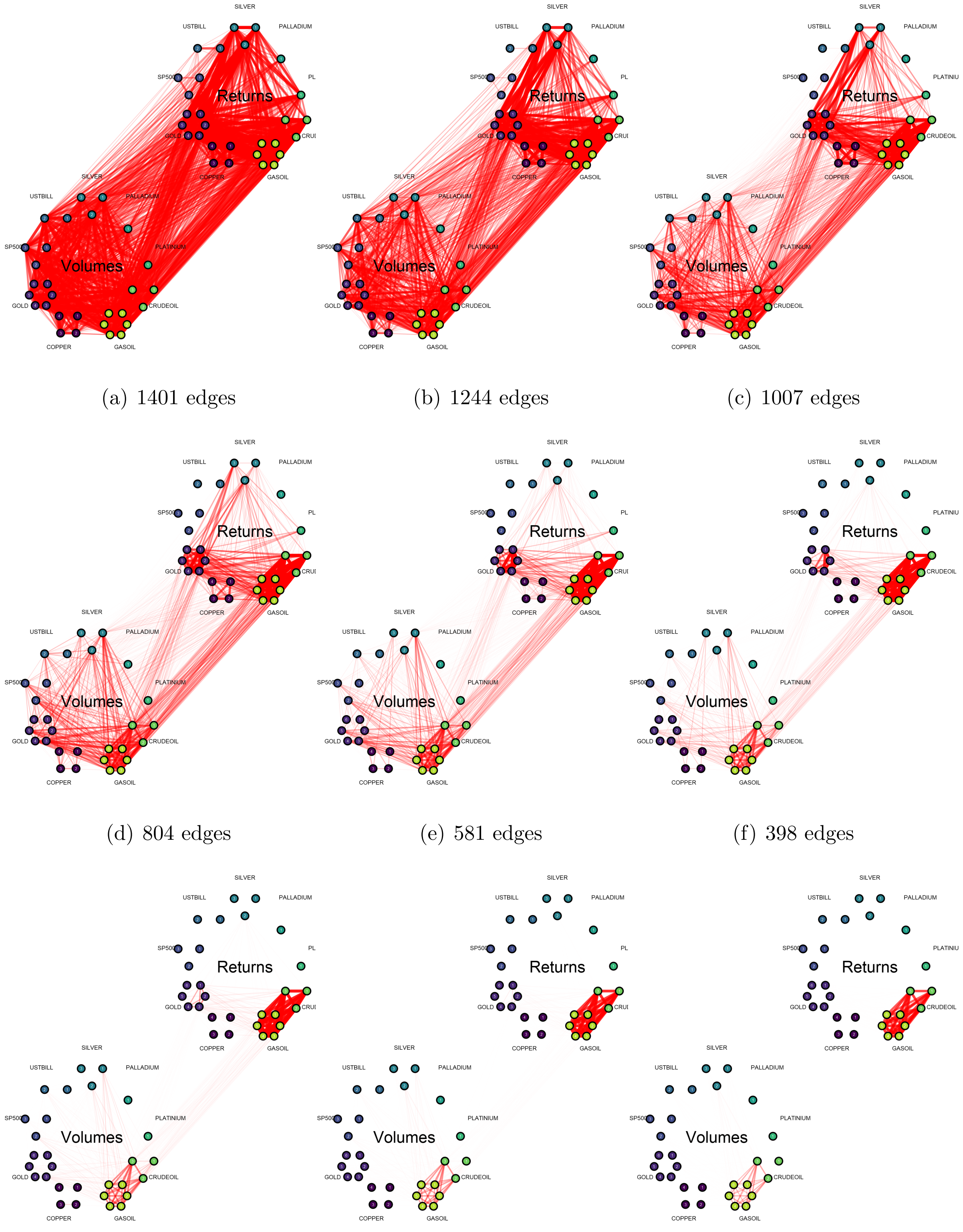
I also proposed a non-parametric method for filling missing values in time series data, based on recurrent neural networks. The simulations show that the method performs well when there are several consecutive missing values, which is the case in my data. A state-space model is fitted on observable data, and the model fill in the blanks. The recurrent neural network makes no a priori on the contemporaneous correlation of the data, which would bias the graph otherwise. That makes it an appropriate filling method in my situation. Still, the solutions are not unique and therefore I end up with a distribution of missing values. I simply chose to keep the median value, but alternative methods for handling that uncertainty are possible. That is left for future research.

Appendix A. Supplementary graphs on original series with missing values

Here, the sparse graphs estimated on raw data (without filling the missing values) are reported. These data have 421,519 observations out of 441,844 in total (7618 days times 58 series). That is still a large sample size. Missing values does not prevent from estimating the graphs, but still that means less observations and possibly some bias in the LLMS estimator. That bias arises because LLMS requires computation of correlation matrices and distances on sub-samples of the data. If these sub-samples have a large proportion of missing values the correlation matrices are known to be biased and so is the selection.

Nonetheless what I observe here, are very similar sparse graphs w.r.t. to the ones estimated on reconstructed data. The analysis of the results in section 4 would be the same on these ones.

Fig. 12. LLMS estimates for a range of penalty



(a) 1401 edges

(b) 1244 edges

(c) 1007 edges

(d) 804 edges

(e) 581 edges

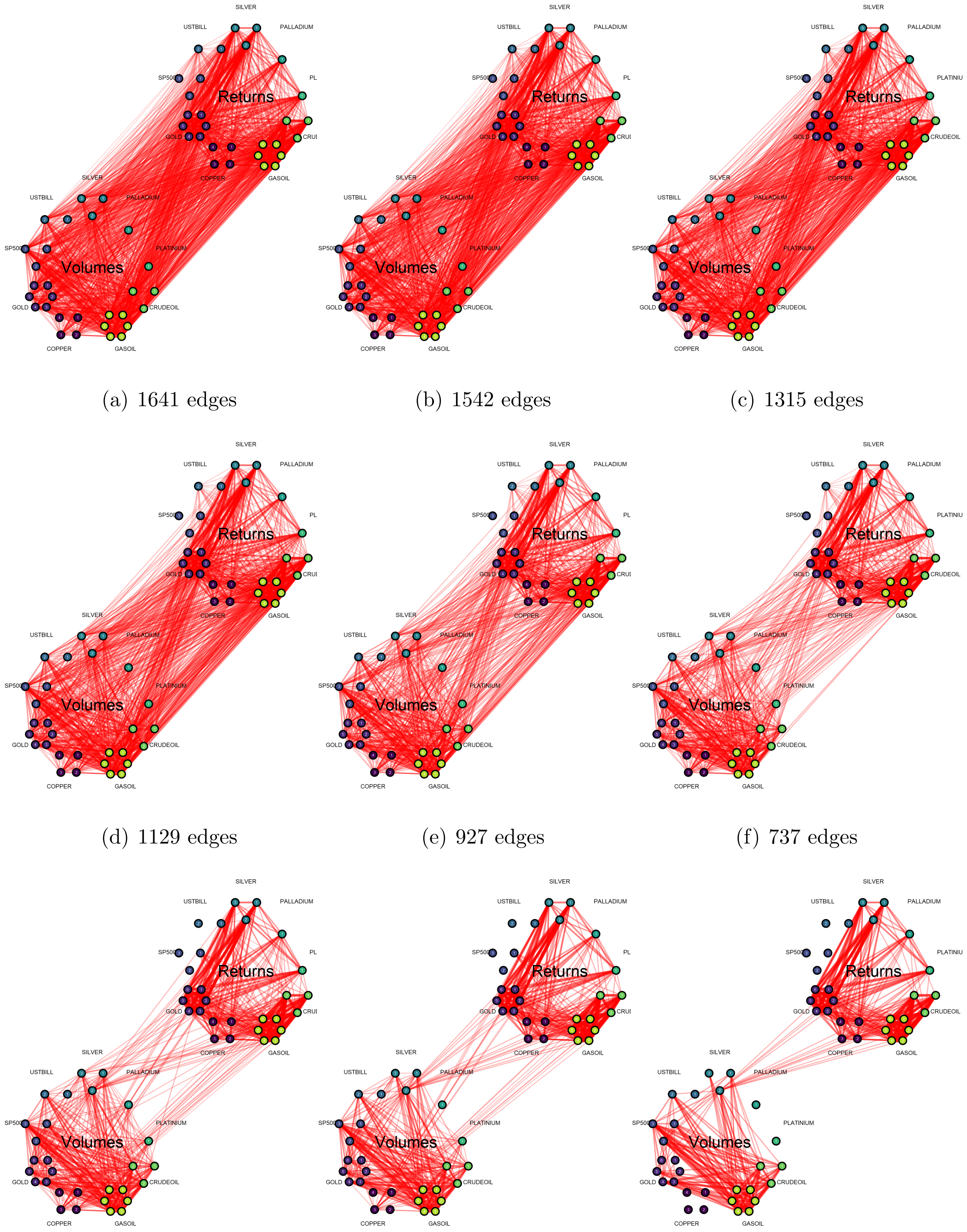
(f) 398 edges

(g) 218 edges

(h) 93 edges

(i) 68 edges

Fig. 13. DAMS estimates for a range of penalty (original series)



(a) 1641 edges

(b) 1542 edges

(c) 1315 edges

(d) 1129 edges

(e) 927 edges

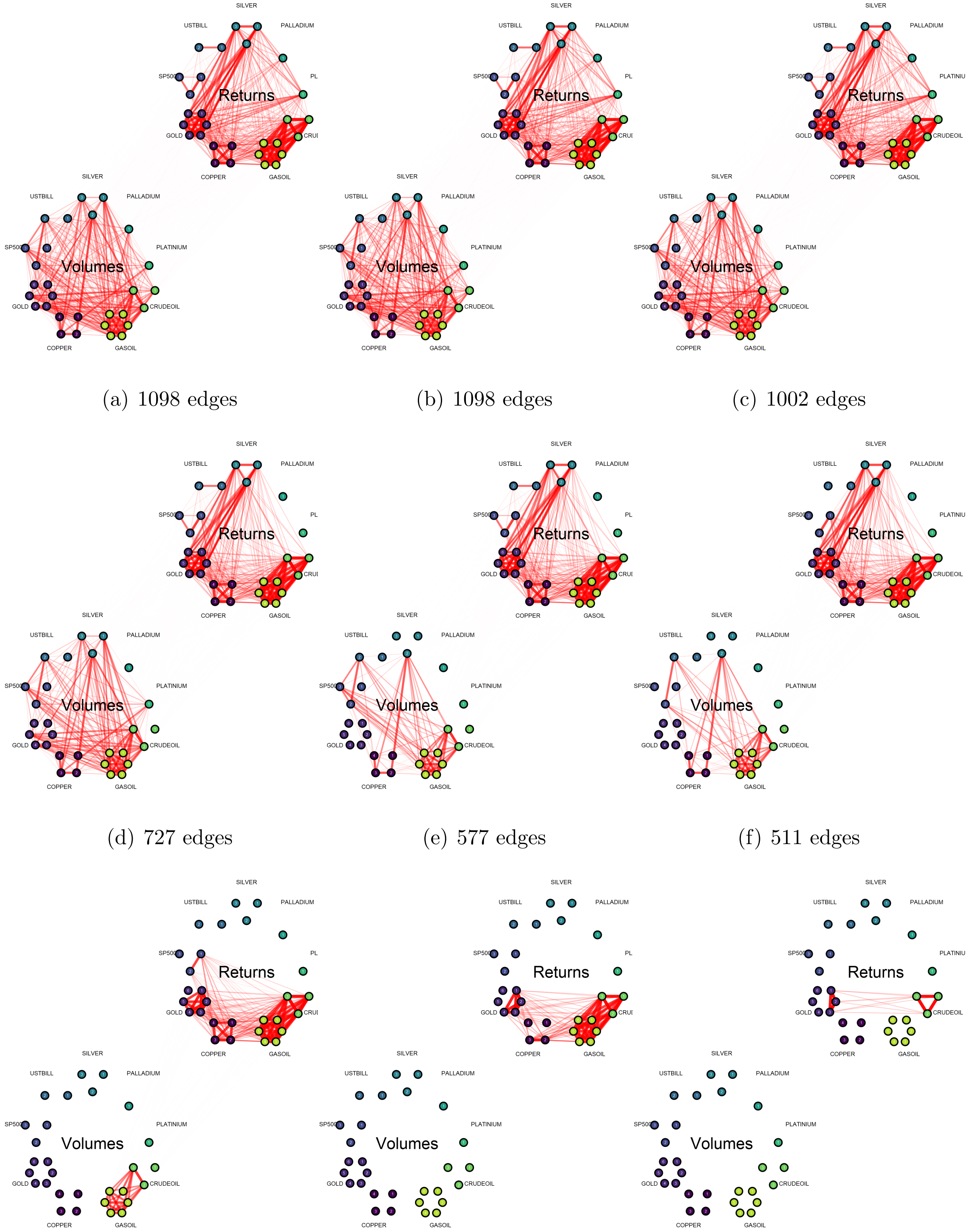
(f) 737 edges

(g) 602 edges

(h) 484 edges

(i) 377 edges

Fig. 14. Graphical LASSO estimates for a range of penalty (original series)



(a) 1098 edges

(b) 1098 edges

(c) 1002 edges

(d) 727 edges

(e) 577 edges

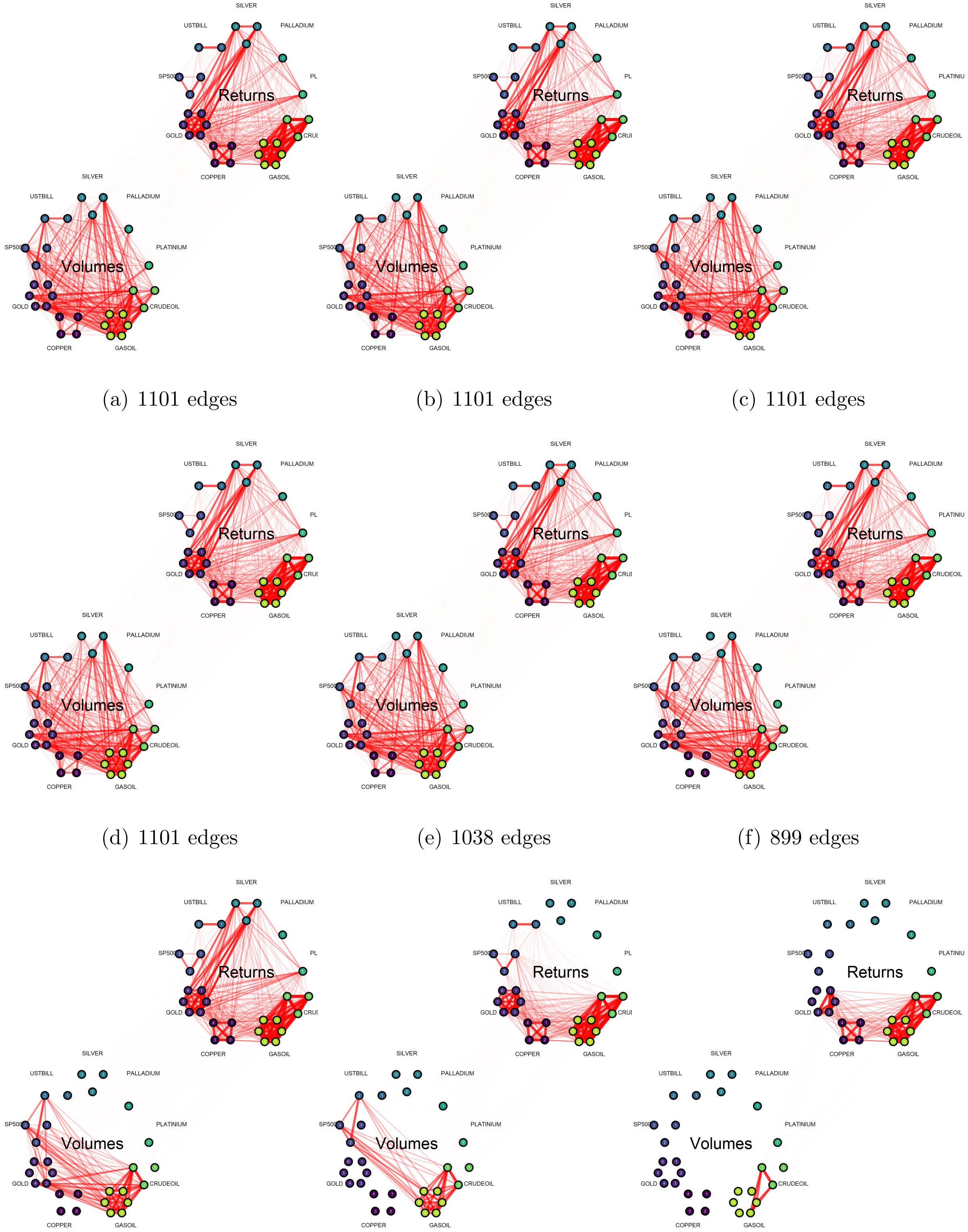
(f) 511 edges

(g) 292 edges

(h) 120 edges

(i) 21 edges

Fig. 15. Nonparanormal estimates for a range of penalty (original series)



(a) 1101 edges

(b) 1101 edges

(c) 1101 edges

(d) 1101 edges

(e) 1038 edges

(f) 899 edges

(g) 661 edges

(h) 398 edges

(i) 173 edges

References

Arbia, G., Bramante, R., Facchinetti, S., and Zappa, D. (2018). Modeling inter-country spatial financial interactions with graphical lasso: An application to sovereign co-risk evaluation. *Regional Science and Urban Economics*, 70:72–79.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research*, 17(1):8162–8181.

Desboulets, L. D. D. (2020). Non-parametric Variable Selection on Non-linear Manifolds. working paper or preprint.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

Lafferty, J., Liu, H., and Wasserman, L. (2012). Sparse nonparametric graphical models. *Statistical Science*, 27(4):519–537.

Millington, T. and Niranjan, M. (2017). Robust portfolio risk minimization using the graphical lasso. pages 863–872.

Murray, J. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–88.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Vink, G., Frank, L., Pannekoek, J., and Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90.

Wang, H. (2010). Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. *Computational Statistics & Data Analysis*, 54(11):2866–2877.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zimmermann, H. G., Grothmann, R., Schäfer, A. M., and Tietz, C. (2005). Modeling large dynamical systems with dynamical consistent neural networks. *New Directions in Statistical Signal Processing*, page 203.