Identifying Small Mean Reverting Portfolios[*]

By Alexandre d'Aspremont[†]

February 26, 2008

Abstract

Given multivariate time series, we study the problem of forming portfolios with maximum mean reversion while constraining the number of assets in these portfolios. We show that it can be formulated as a sparse canonical correlation analysis and study various algorithms to solve the corresponding sparse generalized eigenvalue problems. After discussing penalized parameter estimation procedures, we study the sparsity versus predictability tradeoff and the impact of predictability in various markets.

Keywords: Mean reversion, sparse estimation, convergence trading, momentum trading, covariance selection.

1 Introduction

Mean reversion has received a significant amount of attention as a classic indicator of predictability in financial markets and is sometimes apparent, for example, in equity excess returns over long horizons. While mean reversion is easy to identify in univariate time series, isolating portfolios of assets exhibiting significant mean reversion is a much more complex problem. Classic solutions include cointegration or canonical correlation analysis, which will be discussed in what follows.

One of the key shortcomings of these methods though is that the mean reverting portfolios they identify are dense, i.e. they include every asset in the time series analyzed. For arbitrageurs, this means that exploiting the corresponding statistical arbitrage opportunities involves considerable *transaction costs*. From an econometric point of view, this also impacts the *interpretability* of the resulting portfolio and the significance of the structural relationships it highlights. Finally, optimally mean reverting portfolios often behave like noise and sometimes vary well inside bid-ask spreads, hence do not form meaningful statistical arbitrage opportunities.

Here, we would like to argue that seeking *sparse* portfolios instead, i.e. optimally mean reverting portfolios with a few assets, solves many of these issues at once: fewer assets means less

[*] JEL classification: C61, C82.

[†] ORFE, Princeton University, Princeton, NJ 08544, USA. Email: aspremon@princeton.edu

transaction costs and more interpretable results. In practice, the tradeoff between mean reversion and sparsity is often very favorable. Furthermore, penalizing for sparsity also makes sparse portfolios vary in a wider price range, so the market inefficiencies they highlight are more significant.

Remark that all statements we will make here on mean reversion apply symmetrically to *momentum*. Finding mean reverting portfolios using canonical correlation analysis means minimizing predictability, while searching for portfolios with strong momentum can also be done using canonical correlation analysis, by *maximizing* predictability. The numerical procedures involved are identical.

Mean reversion has of course received a considerable amount of attention in the literature, most authors, such as Fama & French (1988), Poterba & Summers (1988) among many others, using it to model and test for predictability in excess returns. Cointegration techniques (see Engle & Granger (1987), and Alexander (1999) for a survey of applications in finance) are often used to extract mean reverting portfolios from multivariate time series. Early methods relied on a mix of regression and Dickey & Fuller (1979) stationarity tests or Johansen (1988) type tests but it was subsequently discovered that an earlier canonical decomposition technique due to Box & Tiao (1977) could be used to extract cointegrated vectors by solving a generalized eigenvalue problem (see Bewley, Orden, Yang & Fisher (1994) for a more complete discussion).

Several authors then focused on the optimal investment problem when excess returns are mean reverting, with Kim & Omberg (1996) and Campbell & Viceira (1999) or Wachter (2002) for example obtaining closed-form solutions in some particular cases. Liu & Longstaff (2004) also study the optimal investment problem in the presence of a "textbook" finite horizon arbitrage opportunity, modeled as a Brownian bridge, while Jurek & Yang (2006) study this same problem when the arbitrage horizon is indeterminate. Gatev, Goetzmann & Rouwenhorst (2006) studied the performance of pairs trading, using pairs of assets as classic examples of structurally mean-reverting portfolios. Finally, the LTCM meltdown in 1998 focused a lot of attention on the impact of leverage limits and liquidity, see Grossman & Vila (1992) or Xiong (2001) for a discussion.

Sparse estimation techniques in general and the ℓ_1 penalization approach we use here in particular have also received a lot of attention in various forms: variable selection using the LASSO (see Tibshirani (1996)), sparse signal representation using basis pursuit by Chen, Donoho & Saunders (2001), compressed sensing (see Donoho & Tanner (2005) and Candès & Tao (2005)) or covariance selection (see Banerjee, Ghaoui & d'Aspremont (2007)), to cite only a few examples. A recent stream of works on the asymptotic consistency of these procedures can be found in Meinshausen & Yu (2007), Candes & Tao (2007), Banerjee et al. (2007), Yuan & Lin (2007) or Rothman, Bickel, Levina & Zhu (2007) among others.

In this paper, we seek to adapt these results to the problem of estimating sparse (i.e. small) mean reverting portfolios. Suppose that S_{ti} is the value at time t of an asset S_i with $i = 1, \ldots, n$ and $t = 1, \ldots, m$, we form portfolios P_t of these assets with coefficients x_i, and assume they follow an Ornstein-Uhlenbeck process given by:

$$dP_t = \lambda(\bar{P} - P_t)dt + \sigma dZ_t \quad \text{with } P_t = \sum_{i=1}^{n} x_i S_{ti} \tag{1}$$

where Z_t is a standard Brownian motion. Our objective here is to maximize the mean reversion coefficient λ of P_t by adjusting the portfolio weights x_i, under the constraints that $\|x\| = 1$ and that the cardinality of x, *i.e.* the number of nonzero coefficients in x, remains below a given $k > 0$.

Our contribution here is twofold. First, we describe two algorithms for extracting sparse mean reverting portfolios from multivariate time series. One is based on a simple greedy search on the list of assets to include. The other uses semidefinite relaxation techniques to directly get good solutions. Both algorithms use predictability in the sense of Box & Tiao (1977) as a proxy for mean reversion in (1). Second, we show that penalized regression and covariance selection techniques can be used as preprocessing steps to simultaneously stabilize parameter estimation and highlight key dependence relationships in the data. We then study the sparsity versus mean reversion tradeoff in several markets, and examine the impact of portfolio predictability on market efficiency using classic convergence trading strategies.

The paper is organized as follows. In Section 2, we briefly recall the canonical decomposition technique derived in Box & Tiao (1977). In Section 3, we adapt these results and produce two algorithms to extract small mean reverting portfolios from multivariate data sets. In Section 4, we then show how penalized regression and covariance selection techniques can be used as preprocessing tools to both stabilize estimation and isolate key dependence relationships in the time series. Finally, we present some empirical results in Section 5 on U.S. swap rates and foreign exchange markets.

2 Canonical decompositions

We briefly recall below the canonical decomposition technique derived in Box & Tiao (1977). Here, we work in a discrete setting and assume that the asset prices follow a stationary vector autoregressive process with:

$$S_t = S_{t-1}A + Z_t, \tag{2}$$

where S_{t-1} is the lagged portfolio process, $A \in \mathbf{R}^{n\times n}$ and Z_t is a vector of i.i.d. Gaussian noise with zero mean and covariance $\Sigma \in \mathbf{S}^n$, independent of S_{t-1}. Without loss of generality, we can assume that the assets S_t have zero mean. The canonical analysis in Box & Tiao (1977) starts as follows. For simplicity, let us first suppose that $n = 1$ in equation (2), to get:

$$\mathbf{E}[S_t^2] = \mathbf{E}[(S_{t-1}A)^2] + \mathbf{E}\,[Z_t^2],$$

which can be rewritten as $\sigma_t^2 = \sigma_{t-1}^2 + \Sigma$. Box & Tiao (1977) measure the *predictability* of stationary series by:

$$\nu = \frac{\sigma_{t-1}^2}{\sigma_t^2}. \tag{3}$$

The intuition behind this variance ratio is very simple: when it is small the variance of the noise dominates that of S_{t-1} and S_t is almost pure noise, when it is large however, S_{t-1} dominates the noise and S_t is almost perfectly predictable. Throughout the paper, we will use this measure of predictability as a proxy for the mean reversion parameter λ in (1). Consider now a portfolio $P_t = S_t x$ with weights $x \in \mathbf{R}^n$, using (2) we know that $S_t x = S_{t-1}Ax + Z_t x$, and we can measure its predicability as:

$$\nu(x) = \frac{x^T A^T \Gamma A x}{x^T \Gamma x},$$

where Γ is the covariance matrix of S_t. Minimizing predictability is then equivalent to finding the minimum generalized eigenvalue λ solving:

$$\det(\lambda \Gamma - A^T \Gamma A) = 0. \tag{4}$$

Assuming that Γ is positive definite, the portfolio with minimum predictability will be given by $x = \Gamma^{-1/2} z$, where z is the eigenvector corresponding to the smallest eigenvalue of the matrix:

$$\Gamma^{-1/2} A^T \Gamma A \Gamma^{-1/2}. \tag{5}$$

We must now estimate the matrix A. Following Bewley et al. (1994), equation (2) can be written:

$$S_t = \hat{S}_t + \hat{Z}_t,$$

where $\hat{S}_t$ is the least squares estimate of S_t with $\hat{S}_t = S_{t-1}\hat{A}$ and we get:

$$\hat{A} = \left(S_{t-1}^T S_{t-1}\right)^{-1} S_{t-1}^T S_t. \tag{6}$$

The Box & Tiao (1977) procedure then solves for the optimal portfolio by inserting this estimate in the generalized eigenvalue problem above.

Box & Tiao procedure. Using the estimate (6) in (5) and the stationarity of S_t, the Box & Tiao (1977) procedure finds linear combinations of the assets ranked in order of predictability by computing the eigenvectors of the matrix:

$$\left(S_t^T S_t\right)^{-1/2} \left(\hat{S}_t^T \hat{S}_t\right) \left(S_t^T S_t\right)^{-1/2} \tag{7}$$

where $\hat{S}_t$ is the least squares estimate computed above. Figure 1 gives an example of a Box & Tiao (1977) decomposition on U.S. swap rates and shows eight portfolios of swap rates with maturities ranging from one to thirty years, ranked according to predictability. Table 1 shows mean reversion coefficient, volatility and the p-value associated with the mean reversion coefficient. We see that all mean reversion coefficients are significant at the 99% level except for the last portfolio. For this highly mean reverting portfolio, a mean reversion coefficient of 238 implies a half-life of about one day, which explains the lack of significance on daily data.

Bewley et al. (1994) show that the canonical decomposition above and the maximum likelihood decomposition in Johansen (1988) can both be formulated in this manner. We very briefly recall their result below.

Number of swaps:	1	2	3	4	5	6	7	8
Mean reversion	0.58	8.61	16.48	38.59	84.55	174.82	184.83	238.11
P-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51
Volatility	0.21	0.28	0.34	0.14	0.10	0.09	0.07	0.07

Table 1: Summary statistics for canonical U.S. swap portfolios: mean reversion coefficient, volatility and the p-value associated with the mean reversion coefficient for portfolio sizes ranging from one to eight.

Johansen procedure. Following Bewley et al. (1994), the maximum likelihood procedure for estimating cointegrating vectors derived in Johansen (1988) and Johansen (1991) can also be written as a canonical decomposition à la Box & Tiao (1977). Here however, the canonical analysis is performed on the first order differences of the series S_t and their lagged values S_{t-1}. We can rewrite equation (2) as:

$$\Delta S_t = QS_{t-1} + Z_t$$

where $Q = A - \mathbf{I}$. The basis of (potentially) cointegrating portfolios is then found by solving the following generalized eigenvalue problem:

$$\lambda(S_{t-1}^T S_{t-1}) - (S_{t-1}^T \Delta S_t (\Delta S_t^T \Delta S_t)^{-1} \Delta S_t^T S_{t-1}) \tag{8}$$

in the variable $\lambda \in \mathbf{R}$.

3 Sparse decomposition algorithms

In the previous section, we have seen that canonical decompositions can be written as generalized eigenvalue problems of the form:

$$\det(\lambda B - A) = 0 \tag{9}$$

in the variable $\lambda \in \mathbf{R}$, where $A, B \in \mathbf{S}^n$ are symmetric matrices of dimension n. Full generalized eigenvalue decomposition problems are usually solved using a QZ decomposition. Here however, we are only interested in extremal generalized eigenvalues, which can be written in variational form as:

$$\lambda^{\max}(A, B) = \max_{x \in \mathbf{R}^n} \frac{x^T A x}{x^T B x}.$$

In this section, we will seek to maximize this ratio while constraining the cardinality of the (portfolio) coefficient vector x and solve instead:

$$\begin{array}{ll} \text{maximize} & x^T A x / x^T B x \\ \text{subject to} & \mathbf{Card}(x) \leq k \\ & \|x\| = 1, \end{array} \tag{10}$$

where $k > 0$ is a given constant and $\mathbf{Card}(x)$ is the number of nonzero coefficients in x. This will compute a sparse portfolio with maximum predictability (or momentum), a similar problem can

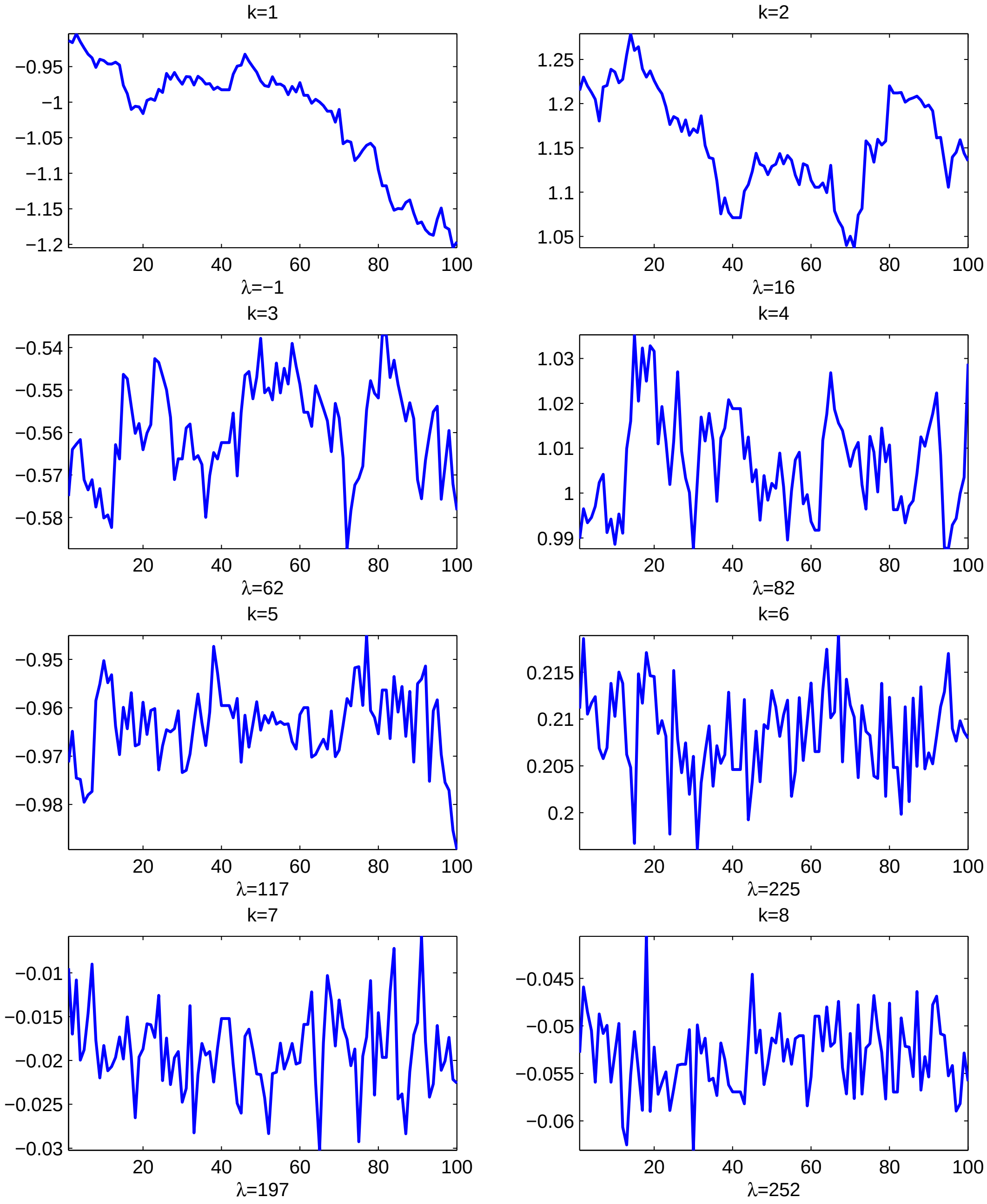

Figure 1: Box-Tiao decomposition on 100 days of U.S. swap rate data (in percent). The eight canonical portfolios of swap rates with maturities ranging from one to thirty years are ranked in decreasing order of predictability. The mean reversion coefficient λ is listed below each plot.

be formed to minimize it (and obtain a sparse portfolio with maximum mean reversion). This is a hard combinatorial problem, in fact, Natarajan (1995) shows that sparse generalized eigenvalue problems are equivalent to subset selection, which is NP-hard. We can't expect to get optimal solutions and we discuss below two efficient techniques to get good approximate solutions.

3.1 Greedy search

Let us call I_k the support of the solution vector x given $k > 0$ in problem (10):

$$I_k = \{i \in [1, n] : \; x_i \neq 0\},$$

by construction $|I_k| \leq k$. We can build approximate solutions to (10) recursively in k. When $k = 1$, we simply find I_1 as:

$$I_1 = \operatorname*{argmax}_{i \in [1,n]} A_{ii}/B_{ii}.$$

Suppose now that we have a good approximate solution with support set I_k given by:

$$x_k = \operatorname*{argmax}_{\{x \in \mathbf{R}^n : \, x_{I_k^c} = 0\}} \frac{x^T A x}{x^T B x},$$

where I_k^c is the complement of the set I_k. This can be solved as a generalized eigenvalue problem of size k. We seek to add one variable with index i_{k+1} to the set I_k to produce the largest increase in predictability by scanning each of the remaining indices in I_k^c. The index i_{k+1} is then given by:

$$i_{k+1} = \operatorname*{argmax}_{i \in I_k^c} \max_{\{x \in \mathbf{R}^n : \, x_{J_i} = 0\}} \frac{x^T A x}{x^T B x}, \quad \text{where } J_i = I_k^c \setminus \{i\},$$

which amounts to solving $(n - k)$ generalized eigenvalue problems of size $k + 1$. We then define:

$$I_{k+1} = I_k \cup \{i_{k+1}\},$$

and repeat the procedure until $k = n$. Naturally, the optimal solutions of problem (10) might not have increasing support sets $I_k \subset I_{k+1}$, hence the solutions found by this recursive algorithm are potentially far from optimal. However, the cost of this method is relatively low: with each iteration costing $O(k^2(n-k))$, the complexity of computing solutions for all target cardinalities k is $O(n^4)$. This recursive procedure can also be repeated forward and backward to improve the quality of the solution.

3.2 Semidefinite relaxation

An alternative to greedy search which has proved very efficient on sparse maximum eigenvalue problems is to derive a convex relaxation of problem (10). In this section, we extend the techniques of d'Aspremont, El Ghaoui, Jordan & Lanckriet (2007) to formulate a semidefinite relaxation for sparse generalized eigenvalue problems in (10):

$$\begin{array}{ll} \text{maximize} & x^T Ax/x^T Bx \\ \text{subject to} & \mathbf{Card}(x) \leq k \\ & \|x\| = 1, \end{array}$$

with variable $x \in \mathbf{R}^n$. As in d'Aspremont et al. (2007), we can form an equivalent program in terms of the matrix $X = xx^T \in \mathbf{S}_n$:

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(AX)/\mathbf{Tr}(BX) \\ \text{subject to} & \mathbf{Card}(X) \leq k^2 \\ & \mathbf{Tr}(X) = 1 \\ & X \succeq 0, \ \mathbf{Rank}(X) = 1, \end{array}$$

in the variable $X \in \mathbf{S}_n$. This program is equivalent to the first one: indeed, if X is a solution to the above problem, then $X \succeq 0$ and $\mathbf{Rank}(X) = 1$ mean that we must have $X = xx^T$, while $\mathbf{Tr}(X) = 1$ implies that $\|x\| = 1$. Finally, if $X = xx^T$ then $\mathbf{Card}(X) \leq k^2$ is equivalent to $\mathbf{Card}(x) \leq k$.

Now, because for any vector $u \in \mathbf{R}^n$, $\mathbf{Card}(u) = q$ implies $\|u\|_1 \leq \sqrt{q}\|u\|_2$, we can replace the nonconvex constraint $\mathbf{Card}(X) \leq k^2$ by a weaker but convex constraint $\mathbf{1}^T|X|\mathbf{1} \leq k$, using the fact that $\|X\|_F = \sqrt{x^T x} = 1$ when $X = xx^T$ and $\mathbf{Tr}(X) = 1$. We then drop the rank constraint to get the following relaxation of (10):

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(AX)/\mathbf{Tr}(BX) \\ \text{subject to} & \mathbf{1}^T|X|\mathbf{1} \leq k \\ & \mathbf{Tr}(X) = 1 \\ & X \succeq 0, \end{array} \tag{11}$$

which is a quasi-convex program in the variable $X \in \mathbf{S}_n$. After the following change of variables:

$$Y = \frac{X}{\mathbf{Tr}(BX)}, \quad z = \frac{1}{\mathbf{Tr}(BX)},$$

and rewrite (11) as:

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(AY) \\ \text{subject to} & \mathbf{1}^T|Y|\mathbf{1} - kz \leq 0 \\ & \mathbf{Tr}(Y) - z = 0 \\ & \mathbf{Tr}(BY) = 1 \\ & Y \succeq 0, \end{array} \tag{12}$$

which is a semidefinite program (SDP) in the variables $Y \in \mathbf{S}_n$ and $z \in \mathbf{R}_+$ and can be solved using standard SDP solvers such as SEDUMI by Sturm (1999) and SDPT3 by Toh, Todd & Tutuncu (1999). The optimal value of problem (12) will be an upper bound on the optimal value of the original problem (10). If the solution matrix Y has rank one, then the relaxation is *tight* and both optimal values are equal. When $\mathbf{Rank}(Y) > 1$ at the optimum in (12), we get an approximate solution to (10) using the rescaled leading eigenvector of the optimal solution matrix Y in (12). The computational complexity of this relaxation is significantly higher than that of the greedy search algorithm in $\S 3.1$. On the other hand, because it is not restricted to increasing sequences of sparse portfolios, the performance of the solutions produced is often higher too. Furthermore, the dual objective value produces an upper bound on suboptimality. Numerical comparisons of both techniques are detailed in Section 5.

4 Parameter estimation

The canonical decomposition procedures detailed in Section 2 all rely on simple estimates of both the covariance matrix Γ in (5) and the parameter matrix A in the vector autoregressive model (6). Of course, both estimates suffer from well-known stability issues and a classic remedy is to penalize the covariance estimation using, for example, a multiple of the norm of Γ. In this section, we would like to argue that using an ℓ_1 penalty term to stabilize the estimation, in a procedure known as covariance selection, simultaneously stabilizes the estimate and helps isolate key idiosyncratic dependencies in the data. In particular, covariance selection clusters the input data in several smaller groups of highly dependent variables among which we can then search for mean reverting (or momentum) portfolios. Covariance selection can then be viewed as a preprocessing step for the sparse canonical decomposition techniques detailed in Section 3. Similarly, penalized regression techniques such as the LASSO by Tibshirani (1996) can be used to produce stable, structured estimates of the matrix parameter A in the VAR model (2).

4.1 Covariance selection

Here, we first seek to estimate the covariance matrix Γ by maximum likelihood. Following Dempster (1972), we penalize the maximum-likelihood estimation to set a certain number of coefficients in the inverse covariance matrix to zero, in a procedure known as *covariance selection*. Zeroes in the inverse covariance matrix correspond to conditionally independent variables in the model and this approach can be used to simultaneously obtain a robust estimate of the covariance matrix while, perhaps more importantly, discovering *structure* in the underlying graphical model (see Lauritzen (1996) for a complete treatment). This tradeoff between log-likelihood of the solution and number of zeroes in its inverse (i.e. model structure) can be formalized in the following problem:

$$\max_{X} \ \log\det X - \mathbf{Tr}(\Sigma X) - \rho\,\mathbf{Card}(X) \tag{13}$$

in the variable $X \in \mathbf{S}_n$, where $\Sigma \in \mathbf{S}_n$ is the sample covariance matrix, $\mathbf{Card}(X)$ is the cardinality of X, i.e. the number of nonzero coefficients in X and $\rho > 0$ is a parameter controlling the tradeoff between likelihood and structure.

Solving the penalized maximum likelihood estimation problem in (13) both improves the stability of this estimation procedure by implicitly reducing the number of parameters and directly highlights structure in the underlying model. Unfortunately, the cardinality penalty makes this problem very hard to solve numerically. One solution developed in d'Aspremont, Banerjee &

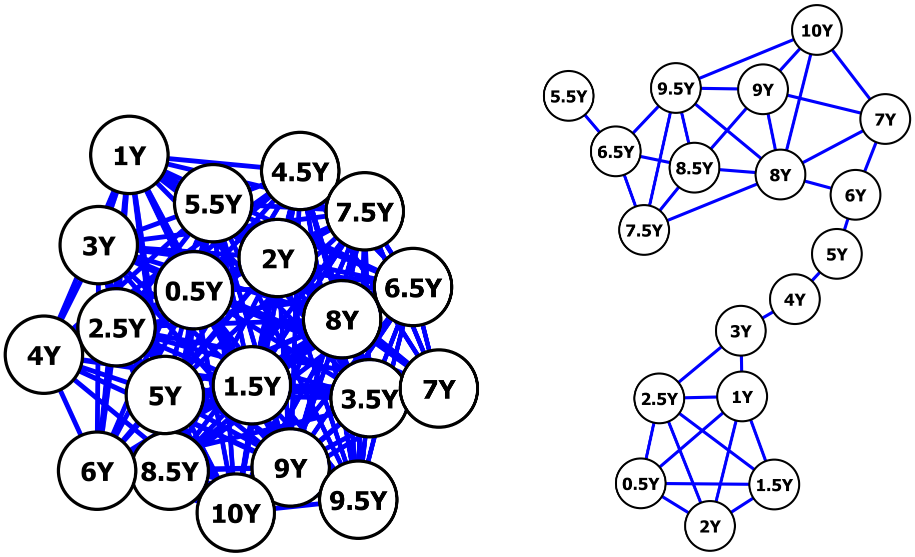

Figure 2: *Left:* conditional dependence network inferred from the pattern of zeros in the inverse swap covariance matrix. *Right:* same plot, using this time the penalized covariance estimate with penalty $\rho = .1$ in the maximum likelihood estimation (14).

El Ghaoui (2006), Banerjee et al. (2007) or Friedman, Hastie & Tibshirani (2007) is to relax the $\mathbf{Card}(X)$ penalty and replace it by the (convex) ℓ_1 norm of the coefficients of X to solve:

$$\max_X \ \log\det X - \mathbf{Tr}(\Sigma X) - \rho \sum_{i,j=1}^{n} |X_{ij}| \tag{14}$$

in the variable $X \in \mathbf{S}^n$. The penalty term involving the sum of absolute values of the entries of X acts as a proxy for the cardinality: the function $\sum_{i,j=1}^{n} |X_{ij}|$ can be seen as the largest convex lower bound on $\mathbf{Card}(X)$ on the hypercube, an argument used by Fazel, Hindi & Boyd (2001) for rank minimization. It is also often used in regression and variable selection procedures, such as the LASSO by Tibshirani (1996). Other permutation invariant estimators have been detailed in Rothman et al. (2007) for example.

In a Gaussian model, zeroes in the inverse covariance matrix point to variables that are conditionally independent, conditioned on all the remaining variables. This has a clear financial interpretation: the inverse covariance matrix reflects independence relationships between the *idiosyncratic* components of asset price dynamics. In Figure 2, we plot the resulting network of dependence, or graphical model for U.S. swap rates. In this graph, variables (nodes) are joined by a link if and only if they are conditionally dependent. We plot the graphical model inferred from the pattern of zeros in the inverse sample swap covariance matrix (left) and the same graph, using this time the penalized covariance estimate in (14) with penalty parameter $\rho = .1$ (right). The graph layout was done using Cytoscape. Notice that in the penalized estimate, rates are clustered by maturity and the graph clearly reveals that swap rates are moving as a curve.

4.2 Estimating structured VAR models

In this section, using similar techniques, we show how to recover a sparse vector autoregressive model from multivariate data.

Endogenous dependence models. Here, we assume that the conditional dependence structure of the assets S_t is purely *endogenous*, i.e. that the noise terms in the vector autoregressive model (2) are i.i.d. with:

$$S_t = S_{t-1}A + Z_t,$$

where $Z_t \sim \mathcal{N}(0, \sigma\mathbf{I})$ for some $\sigma > 0$. In this case, we must have:

$$\Gamma = A^T \Gamma A + \sigma \mathbf{I}$$

since $A^T \otimes A$ has no unit eigenvalue (by stationarity), this means that:

$$\Gamma/\sigma = (\mathbf{I} - A^T \otimes A^T)^{-1}\mathbf{I}$$

where $A \otimes B$ is the Kronecker product of A and B, which implies:

$$A^T A = \mathbf{I} - \sigma\Gamma^{-1}.$$

We can always choose σ small enough so that $\mathbf{I} - \sigma\Gamma^{-1} \succeq 0$. This means that we can directly get A as a matrix square root of $(\mathbf{I} - \sigma\Gamma^{-1})$. Furthermore, if we pick A to be the Cholesky decomposition of $(\mathbf{I} - \sigma\Gamma^{-1})$, and if the graph of Γ is chordal (i.e. has no cycles of length greater than three) then there is a permutation of the variables P such that the Cholesky decomposition of $P\Gamma P^T$, and the upper triangle of $P\Gamma P^T$ have the same pattern of zeroes (see Wermuth (1980) for example). In Figure 4, we plot two dependence networks, one chordal (on the left), one not (on the right). In this case, the structure (pattern of zeroes) of A in the VAR model (6) can be directly inferred from that of the penalized covariance estimate.

Gilbert (1994, §2.4) also shows that if A satisfies $A^T A = \mathbf{I} - \sigma\Gamma^{-1}$ then, barring numerical cancellations in $A^T A$, the graph of Γ^{-1} is the intersection graph of A so:

$$(\Gamma^{-1})_{ij} = 0 \implies A_{ki}A_{kj} = 0, \text{ for all } k = 1, \ldots, n.$$

This means in particular that when the graph of Γ is disconnected, then the graph of A must also be disconnected along the same clusters of variables, i.e. A and Γ have identical block-diagonal structure. In $\S 4.3$, we will use this fact to show that when the graph of Γ is disconnected, optimally mean reverting portfolios must be formed exclusively of assets within a single cluster of this graph.

Exogenous dependence models. In the general case where the noise terms are correlated, with $Z_t \sim \mathcal{N}(0, \Sigma)$ for a certain noise covariance Σ, and the dependence structure is partly exogenous, we need to estimate the parameter matrix A directly from the data. In Section 2, we estimated the matrix A in the vector autoregressive model (2) by regressing S_t on S_{t-1}:

$$\hat{A} = \left(S_{t-1}^T S_{t-1}\right)^{-1} S_{t-1}^T S_t.$$

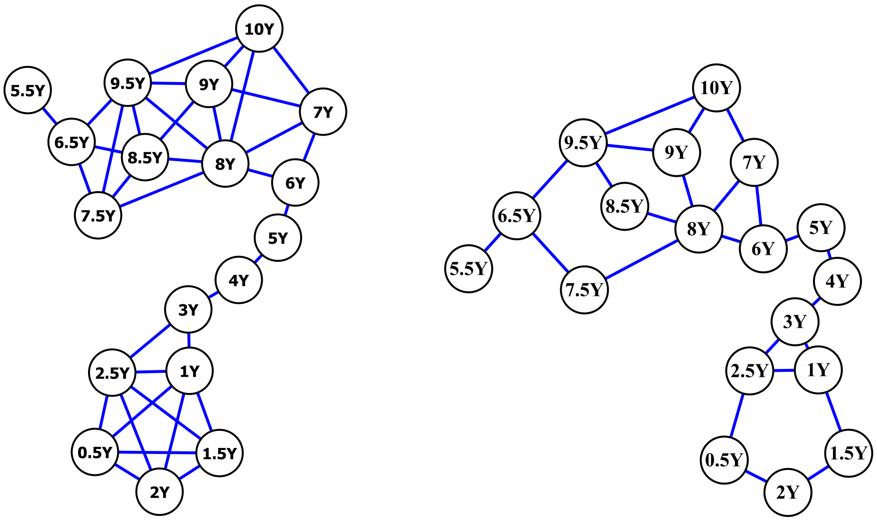

Figure 3: *Left:* a chordal graphical model: no cycles of length greater than three. *Right:* a non-chordal graphical model.

Here too, we can modify this estimation procedure in order to get a sparse model matrix A. Our aim is again to both stabilize the estimation and highlight key dependence relationships between S_t and S_{t-1}. We replace the simple least-squares estimate above by a penalized one. We get the columns of A by solving:

$$a_i = \underset{x}{\operatorname{argmin}} \|S_{it} - S_{t-1}x\|^2 + \gamma\|x\|_1 \tag{15}$$

in the variable $x \in \mathbf{R}^n$, where the parameter $\lambda > 0$ controls sparsity. This is known as the LASSO (see Tibshirani (1996)) and produces sparse least squares estimates.

4.3 Canonical decomposition with penalized estimation

We showed that covariance selection highlights networks of dependence among assets, and that penalized regression could be used to estimate sparse model matrices A. We now show under which conditions these results can be combined to extract information on the support of the canonical portfolios produced by the decompositions in Section 2 from the graph structure of the covariance matrix Γ and of the model matrix A. Because both covariance selection and the lasso are substantially cheaper numerically than the sparse decomposition techniques in Section 3, our goal here is to use these penalized estimation techniques as preprocessing tools to narrow down the range of assets over which we look for mean reversion.

In Section 2, we saw that the Box & Tiao (1977) decomposition for example, could be formed by solving the following generalized eigenvalue problem:

$$\det(\lambda\Gamma - A^T\Gamma A) = 0,$$

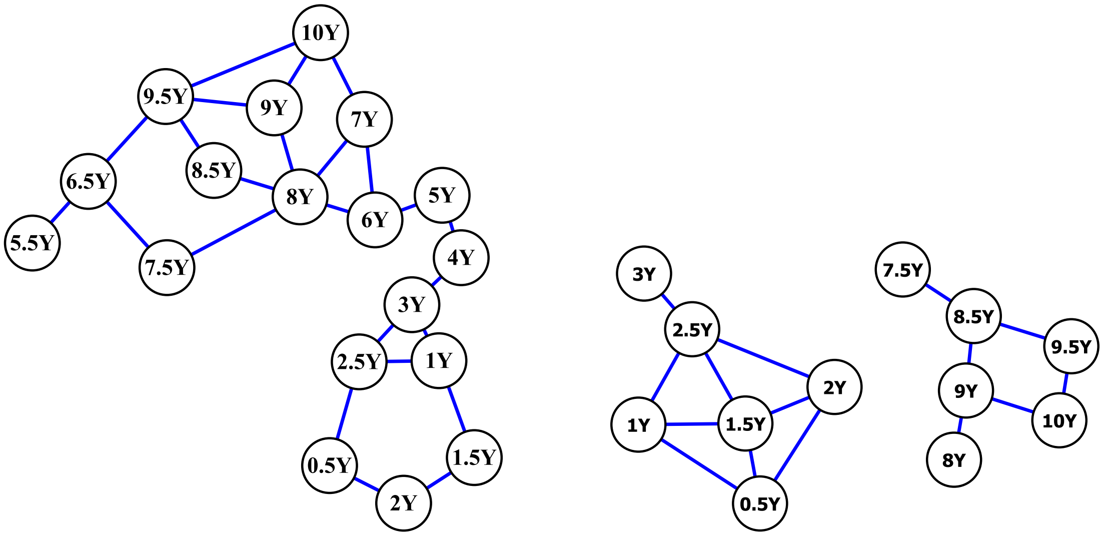

Figure 4: *Left:* a connected graphical model. *Right:* disconnected models.

where Γ is the covariance matrix of the assets S_t and A is the model matrix in (2). Suppose now that our penalized estimates of the matrices Γ and $A^T \Gamma A$ have disconnected graphs with identical clusters, i.e. have the same block diagonal structure, Gilbert (1994, Th. 6.1) shows that the support of the generalized eigenvectors of the pair $\{\Gamma, A^T \Gamma A\}$ must be fully included in one of the clusters of the graph of the inverse covariance Γ^{-1}. In other words, if the graph of the penalized estimate of Γ^{-1} and A are disconnected along the same clusters, then optimally unpredictable (or predictable) portfolios must be formed exclusively of assets in a single cluster.

This suggests a simple procedure for finding small mean reverting portfolios in very large data sets. We first estimate a sparse inverse covariance matrix by solving the covariance selection problem in (14), setting ρ large enough so that the graph of Γ^{-1} is split into sufficiently small clusters. We then check if either the graph is chordal or if penalized estimates of A share some clusters with the graph of Γ^{-1}. After this preprocessing step, we use the algorithms of Section 3 to search these (much smaller) clusters of variables for optimal mean reverting (or momentum) portfolios.

5 Empirical results

In this section, we first compare the performance of the algorithms described in Section 3. We then study the mean reversion versus sparsity tradeoff on various financial instruments. Finally, we test the performance of convergence trading strategies on sparse mean reverting portfolios.

5.1 Numerical performance

In Figure 1 we plotted the result of the Box-Tiao decomposition on U.S. swap rate data (see details below). Each portfolio is a *dense* linear combination of swap rates, ranked in decreasing order of predictability. In Figure 6, we apply the greedy search algorithm detailed in Section 3 to the same data set and plot the *sparse* portfolio processes for each target number of assets. Each subplot of Figure 6 lists the number k of nonzero coefficients of the corresponding portfolio and its mean reversion coefficient λ. Figure 5 then compares the performance of the greedy search algorithm versus the semidefinite relaxation derived in Section 3. On the left, for each algorithm, we plot the mean reversion coefficient λ versus portfolio cardinality (number of nonzero coefficients). We observe on this example that while the semidefinite relaxation does produce better results in some instances, the greedy search is more reliable. Of course, both algorithms recover the same solutions when the target cardinality is set to $k = 1$ or $k = n$. On the right, we plot CPU time (in seconds) as a function of the total number of assets to search. As a quick benchmark, producing 100 sparse mean reverting portfolios for each target cardinality between 1 and 100 took one minute and forty seconds.

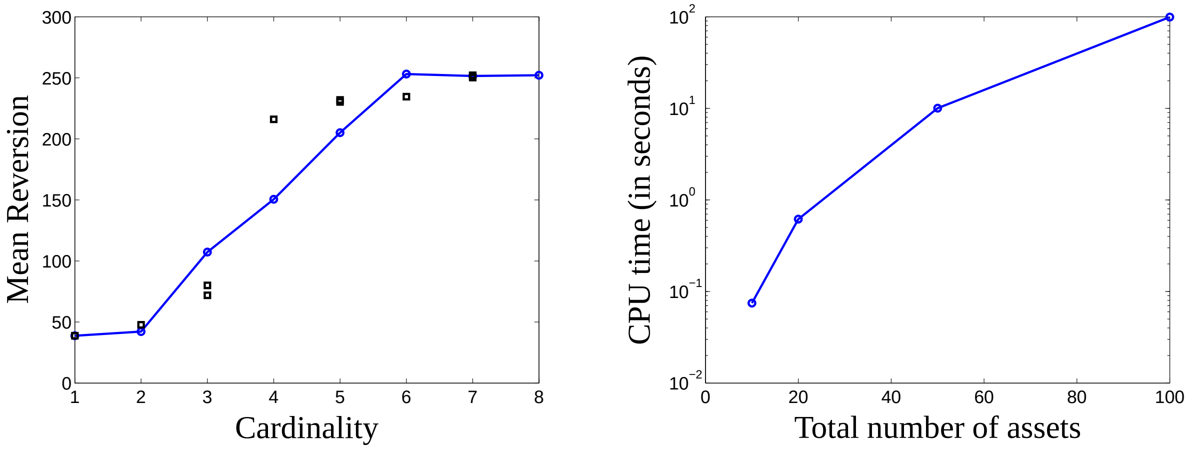

Figure 5: *Left:* Mean reversion coefficient λ versus portfolio cardinality (number of nonzero coefficients) using the greedy search (circles, solid line) and the semidefinite relaxation (squares) algorithms on U.S. swap rate data. *Right:* CPU time (in seconds) versus total number of assets n to compute a full set of sparse portfolios (with cardinality ranging from 1 to n) using the greedy search algorithm.

5.2 Mean reversion versus sparsity

In this section, we study the mean reversion versus sparsity tradeoff on several data sets. We also test the persistence of this mean reversion out of sample.

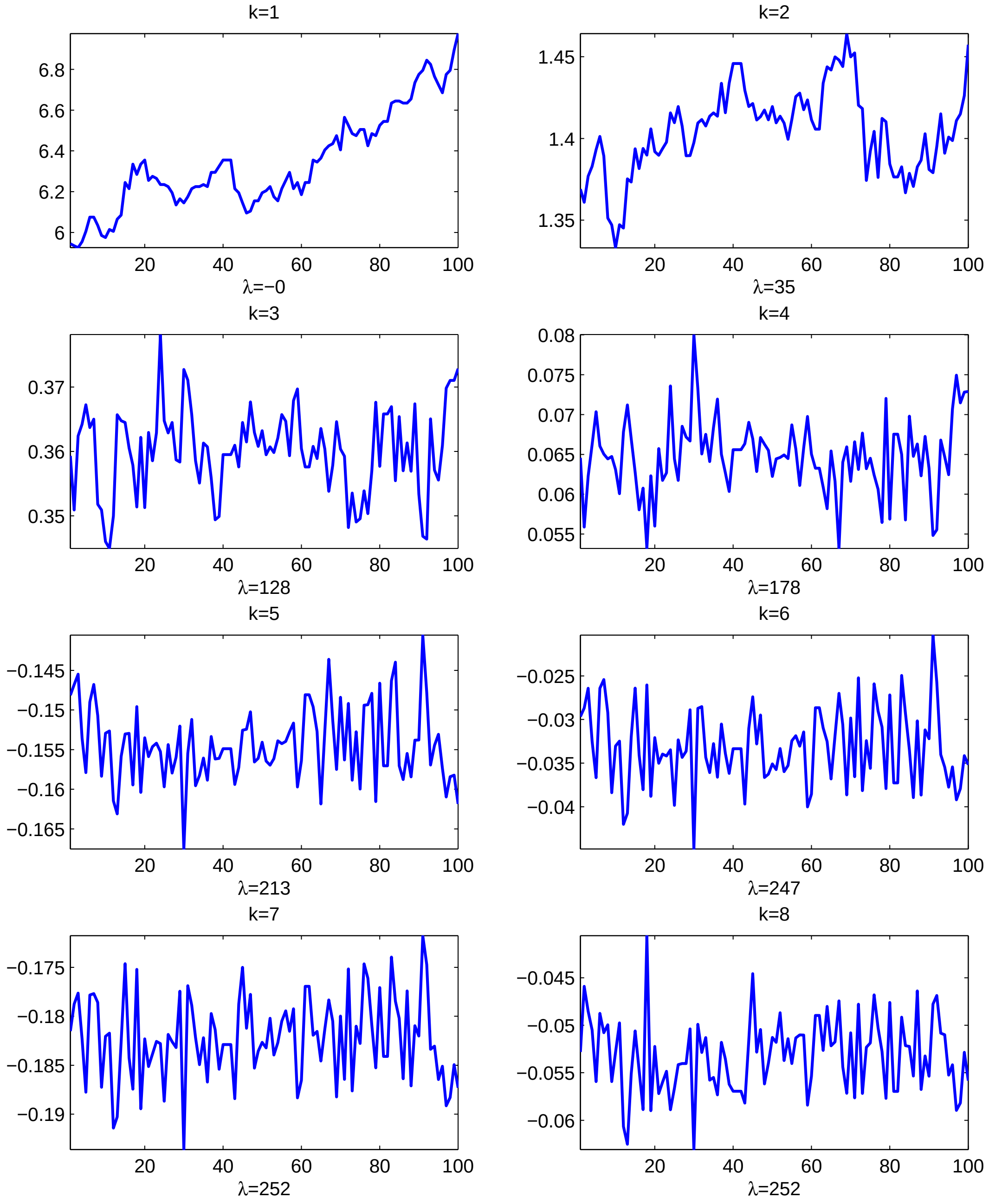

Figure 6: Sparse canonical decomposition on 100 days of U.S. swap rate data (in percent). The number of nonzero coefficients in each portfolio vector is listed as k on top of each subplot, while the mean reversion coefficient λ is listed below each one.

Swap rates. In Figure 7 we compare in and out of sample estimates of the mean reversion versus cardinality tradeoff. We study U.S. swap rate data for maturities 1Y, 2Y, 3Y, 4Y, 5Y, 7Y, 10Y and 30Y from 1998 until 2005. We first use the greedy algorithm of Section 3 to compute optimally mean reverting portfolios of increasing cardinality for time windows of 200 days and repeat the procedure every 50 days. We plot average mean reversion versus cardinality in Figure 7 on the left. We then repeat the procedure, this time computing the (out of sample) mean reversion in the 200 days time window immediately following our sample and also plot average mean reversion versus cardinality. In Figure 7 on the right, we plot the out of sample portfolio price range (spread between min. and max. in basis points) versus cardinality (number of nonzero coefficients) on the same U.S. swap rate data. Table 2 shows the portfolio composition for each target cardinality.

	1	2	3	4	5	6	7	8
1Y	0	0	0	-0.041	-0.037	0.036	-0.013	0.001
2Y	0	0	0	0	0	0	-0.102	0.117
3Y	0	0	-0.288	0.433	0.419	-0.437	0.547	-0.495
4Y	0	-0.714	0.806	-0.803	-0.802	0.809	-0.767	0.702
5Y	1.000	0.700	-0.517	0.408	0.424	-0.389	0.317	-0.427
7Y	0	0	0	0	0	0	0	0.219
10Y	0	0	0	0	0	-0.031	0.025	-0.130
30Y	0	0	0	0	-0.007	0.016	-0.008	0.014

Table 2: Composition of optimal swap portfolios for various target cardinalities.

Foreign exchange rates. We study the following U.S. dollar exchange rates: Argentina, Australia, Brazil, Canada, Chile, China, Colombia, Czech Republic, Egypt, Eurozone, Finland, Hong Kong, Hungary, India, Indonesia, Israel, Japan, Jordan, Kuwait, Latvia, Lithuania, Malaysia, Mexico, Morocco, New Zealand, Norway, Pakistan, Papua NG, Peru, Philippines, Poland, Romania, Russia, Saudi Arabia, Singapore, South Africa, South Korea, Sri Lanka, Switzerland, Taiwan, Thailand, Turkey, United Kingdom, Venezuela, from April 2002 until April 2007. Note that exchange rates are quoted with four digits of accuracy (pip size), with bid-ask spreads around 0.0005 for key rates.

After forming the sample covariance matrix Σ of these rates, we solve the covariance selection problem in (14). This penalized maximum likelihood estimation problem isolates a cluster of 14 rates and we plot the corresponding graph of conditional covariances in Figure 8. For these 14 rates, we then study the impact of penalized estimation of the matrices Γ and A on out of sample mean reversion. In Figure 9, we plot out of sample mean reversion coefficient λ versus portfolio cardinality, on 14 rates selected by covariance selection. The sparse canonical decomposition was performed on both unpenalized estimates and penalized ones. The covariance matrix was estimated by solving the covariance selection problem (14) with $\rho = 0.01$ and the matrix A in (2) was estimated by solving problem (15) with the penalty γ set to zero out 20% of the regression coefficients.

We notice in Figure 9 that penalization has a double impact. First, the fact that sparse portfolios have a higher out of sample mean reversion than dense ones means that penalizing for sparsity helps prediction. Second, penalized estimates of Γ and A also produce higher out of sample mean reversion than unpenalized ones. In Figure 9 on the right, we plot portfolio price range versus cardinality and notice that here too sparse portfolios have a significantly broader range of variation than dense ones.

5.3 Convergence trading

Here, we measure the performance of the convergence trading strategies detailed in the appendix. In Figure 10 we plot average out of sample sharpe ratio versus portfolio cardinality on a 50 days (out of sample) time window immediately following the 100 days over which we estimate the process parameters. Somewhat predictably in the very liquid U.S. swap markets, we notice that while out of sample Sharpe ratios look very promising in frictionless markets, even minuscule transaction costs (a bid-ask spread of 1bp) are sufficient to completely neutralize these market inefficiencies.

6 Conclusion

We have derived two simple algorithms for extracting sparse (i.e. small) mean reverting portfolios from multivariate time series by solving a penalized version of the canonical decomposition technique in Box & Tiao (1977). Empirical results suggest that these small portfolios present a double advantage over their original dense counterparts: sparsity means lower transaction costs and better interpretability, it also improved out-of-sample predictability in the markets studied in Section 5. Several important issues remain open at this point. First, it would be important to show consistency of the variable selection procedure: assuming we know a priori that only a few variables have economic significance (i.e. should appear in the optimal portfolio), can we prove that the sparse canonical decomposition will recover them? Very recent consistency results by Amini & Wainwright (2008) on the sparse principal component analysis relaxation in d'Aspremont et al. (2007) seem to suggest that this is likely, at least for simple models. Second, while the dual of the semidefinite relaxation in (11) provides a bound on suboptimality, we currently have no procedure for deriving simple bounds of this type for the greedy algorithm in Section 3.1.

Acknowledgements

The author would like to thank Marco Cuturi, Guillaume Boulanger and conference participants at the third Cambridge-Princeton conference and the INFORMS 2007 conference in Seattle for helpful comments and discussions. The author would also like to acknowledge support from NSF grant DMS-0625352, ONR grant number N00014-07-1-0150, a Peek junior faculty fellowship and a gift from Google, Inc.

Appendix

In the previous sections, we showed how to extract small mean reverting (or momentum) portfolios from multivariate asset time series. In this section we assume that we have identified such a mean reverting portfolio and model its dynamics given by:

$$dP_t = \lambda(\bar{P} - P_t)dt + \sigma dZ_t, \tag{16}$$

In this section, we detail how to optimally trade these portfolios under various assumptions regarding market friction and risk-management constraints. We begin by quickly recalling results on estimating the Ornstein-Uhlenbeck dynamics in (16).

Estimating Ornstein-Uhlenbeck processes

By explicitly integrating the process P_t in (16) over a time increment Δt we get:

$$P_t = \bar{P} + e^{-\lambda \Delta t}(P_{t-\Delta t} - \bar{P}) + \sigma \int_{t-\Delta t}^{t} e^{\lambda(s-t)} dZ_s, \tag{17}$$

which means that we can estimate λ and σ by simply regressing P_t on P_{t-1} and a constant. With

$$\int_{t-\Delta t}^{t} e^{\lambda(s-t)} dZ_s \sim \sqrt{\frac{1 - e^{-2\lambda\Delta t}}{2\lambda}}\, \mathcal{N}(0,1),$$

we get the following estimators for the parameters of P_t:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{N} \sum_{i=0}^{N} P_t \\
\hat{\lambda} &= -\frac{1}{\Delta t} \log\left(\frac{\sum_{i=1}^{N} (P_t - \hat{\mu})(P_{t-1} - \hat{\mu})}{\sum_{i=1}^{N} (P_t - \hat{\mu})(P_t - \hat{\mu})} \right) \\
\hat{\sigma} &= \sqrt{\frac{2\lambda}{(1 - e^{-2\lambda\Delta t})(N-2)} \sum_{i=1}^{N} \left((P_t - \hat{\mu}) - e^{-\lambda\Delta t}(P_t - \hat{\mu})\right)^2}
\end{aligned}$$

where Δt is the time interval between times t and $t-1$. The expression in (17) also allows us to compute the *half-life* of a market shock on P_t as:

$$\tau = \frac{\log 2}{\lambda}, \tag{18}$$

which is a more intuitive measure of the magnitude of the portfolio's mean reversion.

Utility maximization in frictionless markets

Suppose now that an agent invests in an asset P_t and in a riskless bond B_t following:

$$dB_t = rB_t dt,$$

the wealth W_t of this agent will follow:

$$dW_t = N_t dP_t + (W_t - N_t P_t) r dt.$$

If P_t follows a mean reverting process given by (16), this is also:

$$dW_t = (r(W_t - N_t P_t) + \lambda(\bar{P} - P_t)N_t)dt + N_t \sigma dZ_t.$$

If we write the value function:

$$V(W_t, P_t, t) = \max_{N_t} \mathbf{E}_t \left[e^{-\beta(T-t)} U(W_t) \right],$$

the H.J.B. equation for this problem can be written:

$$\begin{aligned} \beta V \quad = \quad & \max_{N_t} \frac{\partial V}{\partial P}\lambda(\bar{P}_t - P_t) + \frac{\partial V}{\partial W}(r(W_t - N_t P_t) + \lambda(\bar{P} - P_t)N_t) + \frac{\partial V}{\partial t} \\ & + \frac{1}{2}\frac{\partial^2 V}{\partial P^2}\sigma^2 + \frac{1}{2}\frac{\partial^2 V}{\partial P \partial W} N_t \sigma^2 + \frac{1}{2}\frac{\partial^2 V}{\partial W^2} N_t^2 \sigma^2 \end{aligned}$$

Maximizing in N_t yields the following expression for the number of shares in the optimal portfolio:

$$N_t = \frac{\partial V / \partial W}{\partial^2 V / \partial W^2 \sigma^2}(\lambda(\bar{P} - P_t) - rP_t) - \frac{\partial^2 V / \partial P \partial W}{\partial^2 V / \partial W^2} \tag{19}$$

Jurek & Yang (2006) solve this equation explicitly for $U(x) = \log x$ and $U(x) = x^{1-\gamma}/(1-\gamma)$ and we recover in particular the classic expression:

$$N_t = \left(\frac{\lambda(\bar{P} - P_t) - rP_t}{\sigma^2} \right) W_t,$$

in the log-utility case.

Leverage constraints

Suppose now that the portfolio is subject to fund withdrawals so that the total wealth evolves according to:

$$dW = d\Pi + dF$$

where $d\Pi = N_t dP_t + (W_t - N_t P_t) r dt$ and dF represents fund flows, with:

$$dF = f d\Pi + \sigma_f dZ_t^{(2)}$$

where $Z_t^{(2)}$ is a Brownian motion (independent of Z_t). Jurek & Yang (2006) show that the optimal portfolio allocation can also be computed explicitly in the presence of fund flows, with:

$$N_t = \left(\frac{\lambda(\bar{P} - P_t) - rP_t}{\sigma^2}\right) \frac{1}{(1+f)} W_t = L_t W_t,$$

in the log-utility case. Note that the constant f can also be interpreted in terms of leverage limits. In steady state, we have:

$$P_t \sim \mathcal{N}\left(\bar{P}, \frac{\sigma^2}{2\lambda}\right)$$

which means that the leverage L_t itself is normally distributed. If we assume for simplicity that $\bar{P} = 0$, given the fund flow parameter f, the leverage will remain below the level M given by:

$$M = \frac{\alpha(\lambda + r)}{(1+f)\sigma\sqrt{2\lambda}} \tag{20}$$

with confidence level $N(\alpha)$, where $N(x)$ is the Gaussian CDF. The bound on leverage M can thus be seen as an alternate way of identifying or specifying the fund flow constant f in order to manage capital outflow risks.

References

Alexander, C. (1999), ‘Optimal hedging using cointegration’, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **357**(1758), 2039–2058.

Amini, A. & Wainwright, M. (2008), ‘High dimensional analysis of semidefinite relaxations for sparse principal component analysis’, *Tech report, Statistics Dept., U.C. Berkeley* .

Banerjee, O., Ghaoui, L. E. & d’Aspremont, A. (2007), ‘Model selection through sparse maximum likelihood estimation’, ICML06. To appear in Journal of Machine Learning Research.

Bewley, R., Orden, D., Yang, M. & Fisher, L. (1994), ‘Comparison of Box-Tiao and Johansen Canonical Estimators of Cointegrating Vectors in VEC (1) Models’, *Journal of Econometrics* **64**, 3–27.

Box, G. E. & Tiao, G. C. (1977), ‘A canonical analysis of multiple time series’, *Biometrika* **64**(2), 355.

Campbell, J. & Viceira, L. (1999), ‘Consumption and Portfolio Decisions When Expected Returns Are Time Varying’, *The Quarterly Journal of Economics* **114**(2), 433–495.

Candès, E. J. & Tao, T. (2005), ‘Decoding by linear programming’, *Information Theory, IEEE Transactions on* **51**(12), 4203–4215.

Candes, E. & Tao, T. (2007), ‘The Dantzig selector: statistical estimation when$ p$ is much larger than$ n$’, *To appear in Annals of Statistics* .

Chen, S., Donoho, D. & Saunders, M. (2001), ‘Atomic decomposition by basis pursuit.’, *SIAM Review* **43**(1), 129–159.

d'Aspremont, A., Banerjee, O. & El Ghaoui, L. (2006), 'First-order methods for sparse covariance selection', *To appear in SIAM Journal on Matrix Analysis and Applications* .

d'Aspremont, A., El Ghaoui, L., Jordan, M. & Lanckriet, G. R. G. (2007), 'A direct formulation for sparse PCA using semidefinite programming', *SIAM Review* **49**(3), 434–448.

Dempster, A. (1972), 'Covariance selection', *Biometrics* **28**, 157–175.

Dickey, D. & Fuller, W. (1979), 'Distribution of the Estimators for Autoregressive Time Series With a Unit Root', *Journal of the American Statistical Association* **74**(366), 427–431.

Donoho, D. L. & Tanner, J. (2005), 'Sparse nonnegative solutions of underdetermined linear equations by linear programming', *Proc. of the National Academy of Sciences* **102**(27), 9446–9451.

Engle, R. & Granger, C. (1987), 'Cointegration and error correction: representation, estimation and testing', *Econometrica* **55**(2), 251–276.

Fama, E. & French, K. (1988), 'Permanent and Temporary Components of Stock Prices', *The Journal of Political Economy* **96**(2), 246–273.

Fazel, M., Hindi, H. & Boyd, S. (2001), 'A rank minimization heuristic with application to minimum order system approximation', *Proceedings American Control Conference* **6**, 4734–4739.

Friedman, J., Hastie, T. & Tibshirani, R. (2007), 'Sparse inverse covariance estimation with the lasso', *Working paper* .

Gatev, E., Goetzmann, W. & Rouwenhorst, K. (2006), 'Pairs Trading: Performance of a Relative-Value Arbitrage Rule', *Review of Financial Studies* **19**(3), 797.

Gilbert, J. (1994), 'Predicting Structure in Sparse Matrix Computations', *SIAM Journal on Matrix Analysis and Applications* **15**(1), 62–79.

Grossman, S. & Vila, J. (1992), 'Optimal Dynamic Trading with Leverage Constraints', *The Journal of Financial and Quantitative Analysis* **27**(2), 151–168.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**(2/3), 231–254.

Johansen, S. (1991), 'Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models', *Econometrica* **59**(6), 1551–1580.

Jurek, J. & Yang, H. (2006), Dynamic portfolio selection in arbitrage, Technical report, Working Paper, Harvard Business School.

Kim, T. & Omberg, E. (1996), 'Dynamic Nonmyopic Portfolio Behavior', *The Review of Financial Studies* **9**(1), 141–161.

Lauritzen, S. (1996), 'Graphical Models'.

Liu, J. & Longstaff, F. (2004), 'Losing Money on Arbitrage: Optimal Dynamic Portfolio Choice in Markets with Arbitrage Opportunities', *Review of Financial Studies* **17**(3).

Meinshausen, N. & Yu, B. (2007), Lasso-type recovery of sparse representations for highdimensional data, Technical report, To appear in Annals of Statistics.

Natarajan, B. K. (1995), 'Sparse approximate solutions to linear systems', *SIAM J. Comput.* **24**(2), 227–234.

Poterba, J. M. & Summers, L. H. (1988), 'Mean reversion in stock prices: Evidence and implications', *Journal of Financial Economics* **22**(1), 27–59.

Rothman, A., Bickel, P., Levina, E. & Zhu, J. (2007), 'Sparse permutation invariant covariance estimation', *Technical report 467, Dept. of Statistics, Univ. of Michigan* .

Sturm, J. (1999), 'Using SEDUMI 1.0x, a MATLAB toolbox for optimization over symmetric cones', *Optimization Methods and Software* **11**, 625–653.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the LASSO', *Journal of the Royal statistical society, series B* **58**(1), 267–288.

Toh, K. C., Todd, M. J. & Tutuncu, R. H. (1999), 'SDPT3 – a MATLAB software package for semidefinite programming', *Optimization Methods and Software* **11**, 545–581.

Wachter, J. (2002), 'Portfolio and Consumption Decisions under Mean-Reverting Returns: An Exact Solution for Complete Markets', *The Journal of Financial and Quantitative Analysis* **37**(1), 63–91.

Wermuth, N. (1980), 'Linear Recursive Equations, Covariance Selection, and Path Analysis', *Journal of the American Statistical Association* **75**(372), 963–972.

Xiong, W. (2001), 'Convergence trading with wealth effects: an amplification mechanism in financial markets', *Journal of Financial Economics* **62**(2), 247–292.

Yuan, M. & Lin, Y. (2007), 'Model selection and estimation in the Gaussian graphical model', *Biometrika* **94**(1), 19.

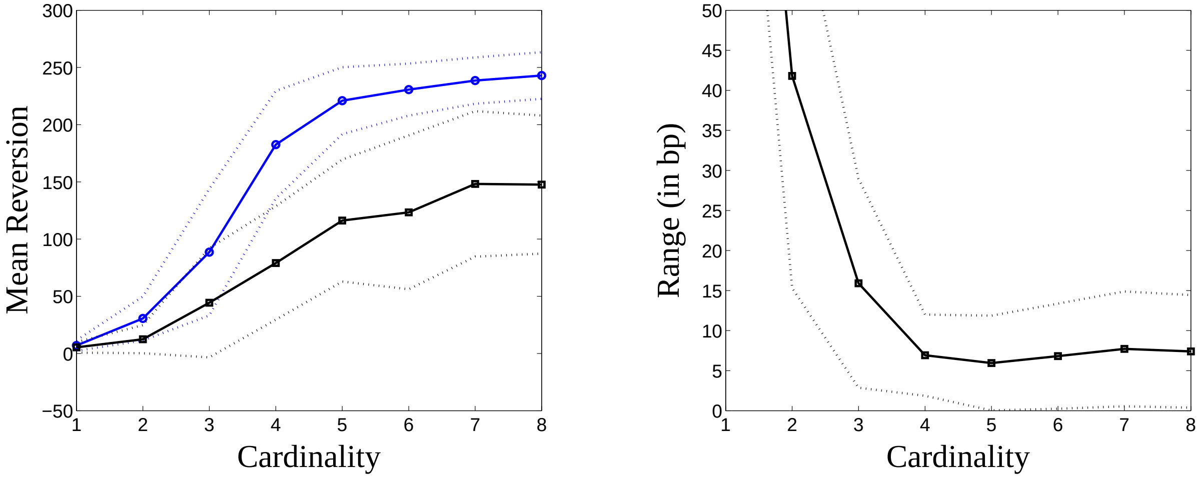

Figure 7: *Left:* mean reversion coefficient λ versus portfolio cardinality (number of nonzero coefficients), in sample (blue circles) and out of sample (black squares) on U.S. swaps. *Right:* out of sample portfolio price range (in basis points) versus cardinality (number of nonzero coefficients) on U.S. swap rate data. The dashed lines are at plus and minus one standard deviation.

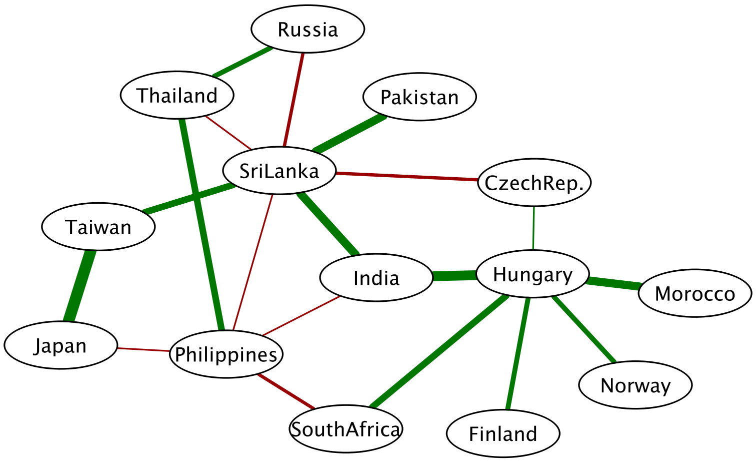

Figure 8: Graph of conditional covariance among a cluster of U.S. dollar exchange rates. Positive dependencies are plotted as green links, negative ones in red, while the thickness reflects the magnitude of the covariance.

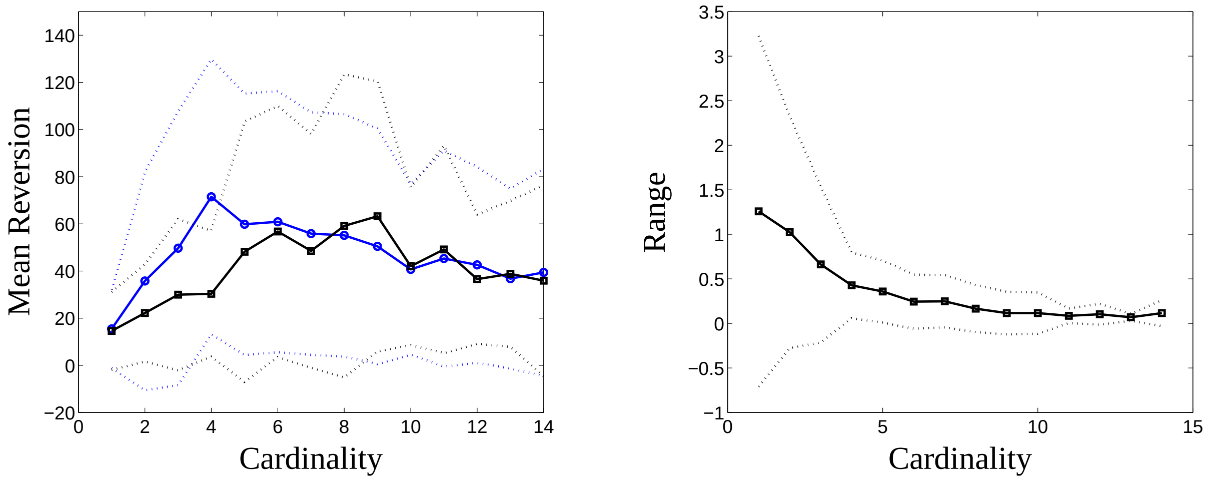

Figure 9: *Left:* out of sample mean reversion coefficient versus portfolio cardinality (number of nonzero coefficients), on 14 U.S. dollar exchange rates clustered by covariance selection. The sparse canonical decomposition was performed on both unpenalized estimates (black squares) and penalized ones (blue circles). *Right:* out of sample portfolio price range (in percent) versus cardinality. The dashed lines are at plus and minus one standard deviation.

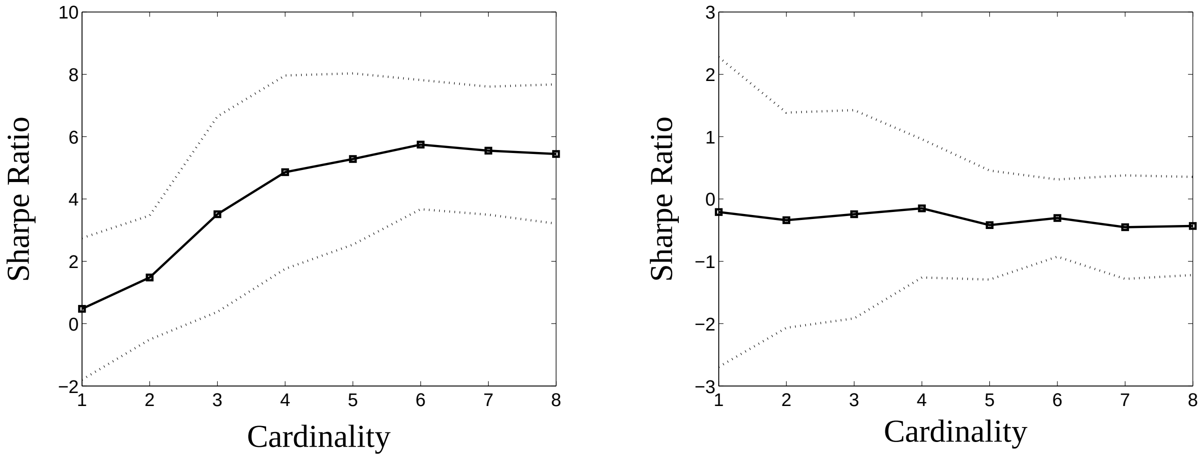

Figure 10: *Left:* average out of sample sharpe ratio versus portfolio cardinality on U.S. swaps. *Right:* idem, with a bid-ask spread of 1bp. The dashed lines are at plus and minus one standard deviation.