6-2019

Biclustering via mixtures of regression models

Raja VELU

Zhaoque ZHOU

Chyng Wen TEE
Singapore Management University, cwtee@smu.edu.sg

Biclustering via Mixtures of Regression Models

Raja Velu[1(✉)], Zhaoque Zhou[1], and Chyng Wen Tee[2]

[1] Whitman School of Management, Syracuse University,
Syracuse, NY 13244, USA
{rpvelu,zzhou37}@syr.edu
[2] Lee Kong Chian School of Business, Singapore Management University,
Singapore 178899, Singapore
cwtee@smu.edu.sg

Abstract. Biclustering of observations and the variables is of interest in many scientific disciplines; In a single set of data matrix it is handled through the singular value decomposition. Here we deal with two sets of variables: Response and predictor sets. We model the joint relationship via regression models and then apply SVD on the coefficient matrix. The sparseness condition is introduced via Group Lasso; the approach discussed here is quite general and is illustrated with an example from Finance.

Keywords: Multivariate regression · Singular value decomposition · Dimension reduction · Mixture models

1 Introduction

In the area of data mining for high-dimensional data with 'm' units and 'n' variables, where 'm' and 'n' are very large, there is a great deal of interest to reduce the dimensions on both sides. For the reduction of large number of variables to a smaller number, the techniques based on Principal Component Analysis (PCA) are normally used. The interpretation of these components is usually improved by the sparseness constraints such as LASSO or RIDGE. On the other hand, the reduction in the large number of units is handled by clustering techniques or through finite mixtures models. In the former, the focus is on correlation between the variables, and in the latter the focus is on the distance between units with appropriate standardization of the variables. In this paper, we want to study the relationship between two sets of variables Y_t and X_t, which are of dimensions 'm' and 'n' collected over 'T' time periods via multivariate regression model and we want to reduce the dimension of the $(m \times n)$ coefficient matrix on both units and variables. Biclustering methods generally focus on a single data matrix; here we focus on the estimated coefficient matrix that relates Y_t to X_t, that represents both the data matrix Y and X. But this has to be properly weighted in for the estimation errors.

In Sect. 2, we introduce the model, review the methodology of clustering of regression models. The following section will cover the estimation methods and some approximate solutions for biclustering. In Sect. 4, we define the problem of biclustering of the regression models and evaluate the procedures using both simulated and real data. The final section will outline the future work.

2 Clustering of Regression Models

In microarray experiments, the objectives are twofold: to detect differentially expressed genes and to group genes with similar expression. It is believed that genes with different expression can provide disease-specific markers and co-expressed genes can contribute to understanding the regulatory network aspect of the gene expression (see Qin and Self [5]). In finance, there is a great deal of interest in assessing the commonality in returns among stocks and how this commonality can be related to commonality in order flow (see Hasbrouck and Seppi [3]). This research has led to fundamental questions related to diversification—a central tenet of Markowitz's portfolio theory. The basic regression model can be written as:

$$\underset{T\times 1}{Y_i} = \underset{T\times n}{X_i} \cdot \underset{n\times 1}{c_i} + e_i, \quad i = 1, \ldots, m, \tag{1}$$

where the times series of responses (returns) on unit i are given in Y_i and the corresponding (order flow) variables related to ith response is given in X_i.

If e_i's are assumed to be correlated over 'i', then the above model set-up is called seemingly unrelated regression equations (SURE). The focus on the clustering methods is on in-grouping 'c_i' into gene-based groups, or in the finance context into trading-pattern based or into industry groups.

Finite Mixture Regression Models: We will assume that there are 'K' groups and each set of models in (1) can come from one of these groups. The random indicator is given by S_i is distributed with unknown probability distribution $\eta = (\eta_1, \ldots, \eta_K)$. The unknown parameters are given in $\theta = (\eta_1, \ldots, \eta_K, c_1, \ldots, c_K, \Sigma_1, \ldots, \Sigma_K)$. The marginal distribution of Y_i given X_i and θ can be written as:

$$\begin{aligned} f(y_i \mid X_i, \theta) &= \sum_{k=1}^{K} f(y_i \mid X_i, S_i, \theta) f(S_i = k \mid \theta) \\ &= \sum_{k=1}^{K} \eta_k f(y_i, X_i c_k, \Sigma_k), \end{aligned} \tag{2}$$

as stated in Frühwirth-Schalter [2, p. 243]. It is important to note that the regression coefficients are identifiable if and only if the number of clusters is less than the number of distinct $(n-1)$ dimensional hyperplanes generated by the covariates. If the covariates show not much variability, there may be identifiability problems which indicates that the groups may not be that distinctive.

Because the cluster membership is not known a priori, treating them as missing information and then using the EM algorithm is what is typically followed. The initial cluster centers are formed from the least squares estimate of c's and grouping them via empirical clustering procedures such as K-means clustering. The number of clusters, the choice of 'K' is made through some criterion of reproducibility over two or more clustering samples. The following BIC criterion is suggested for the selection of 'K':

$$\text{BIC}_K = 2\text{loglikelihood} - n \ln T. \tag{3}$$

Qin and Self [5] also suggest a measure for the stability of cluster centers:

$$\text{BMV}_K = \max_{k=1\ldots,K}(\text{volume}(\hat{\Sigma}_k)), \tag{4}$$

where BMV is the bootstrapped maximum value and the volume is measured by largest eigenvalue. The joint use of BIC and BMV will lead to a selection of 'K' value that has large BIC and a small BMV.

The clustering of regression models is fit by applying the EM algorithm as follows: stack the data as $Y = (Y_1', \ldots, Y_m')'$ and $X = (X_1', \ldots, X_m')'$, $S = (S_1, \ldots, S_m)'$ and $\theta = (c_1', \ldots, c_K', \eta_1, \ldots, \eta_K)$ be the stacked parameter vector. The E-step finds the expectation of unknown variables, such as cluster membership and the M-step maximizes the likelihood to obtain parameter estimates. These result in the following steps:

$$\begin{aligned}
\hat{s}_{ik} &= \frac{\eta_k \Pr(y_i \mid s_{ik} = 1; \theta)}{\sum_{k=1}^{K} \eta_k \Pr(y_i \mid s_{ik} = 1; \theta)} \\
\hat{\eta}_k &= \sum_{i=1}^{m} \hat{s}_{ik} \\
\hat{c}_k &= \left(\sum_{i=1}^{m} \hat{s}_{ik} X_i' X_i \right)^{-1} \sum_{i=1}^{m} \hat{s}_{ik} X_i' Y_i \\
\hat{\sigma}_k^2 &= \sum_{i=1}^{m} \hat{s}_{ik} (Y_i - X_i \hat{c}_k)'(Y_i - X_i \hat{c}_k) / \sum_{i=1}^{m} T \hat{s}_{ik}.
\end{aligned} \tag{5}$$

The convergence of these estimates depend on the choice of initial values.

From the structure of the estimates of 'c' coefficients in (5), it is useful to note that the cluster 'c_k' coefficients are weighted average of the models entertained with weights being assigned by the chance of their belongings to the cluster membership. Note

$$\hat{c}_k = (\sum_{i=1}^{m} w_i)^{-1} (\sum_{i=1}^{m} w_i \tilde{c}_i) \tag{6}$$

where $w_i = \hat{s}_{ik} X_i^{'} X_i$ and $\tilde{c}_i = (X_i^{'} X_i)^{-1} X_i^{'} Y_i$, the least squares estimate.If the design matrix X_i's are same, then the above simplifies to,

$$\hat{c}_k = \sum_{i=1}^{m} \hat{s}_{ik} \tilde{c}_i \tag{7}$$

which implies the likelihood equations based on the data can be replaced by the likelihood equations using the least squares estimates and their distributions.

3 Biclustering of Observations and Variables

In the last section the focus was on clustering subjects but clustering of variables is also very useful in practice. The similarity in variables is captured through PCA. The concept of biclustering was discussed earlier by Rao [6] who advocated computing the singular value decomposition (SVD) of the $m \times n$ data matrix Y where 'm' is the number of subjects and 'n' is the number of variables and also using the SVD of Y'. The biclustering procedure simultaneously clusters both the subjects and the variables thus providing a way to be selective of both subjects and the variables. This is quite useful if any investments in follow-up studies are expensive.

The sparse singular value decomposition (SSVD) is proposed as an exploratory tool for biclustering in Lee, Shen, Huang and Marron [4]. The requirement is that for a '$m \times n$' data matrix, Y, use the following penalized sum-of-squares criterion.

$$\|Y - suv'\|_F^2 + \lambda_u P_1(su) + \lambda_v P_2(sv) \tag{8}$$

where the first term is the squared Frobenius norm, P_1 and P_2 are sparse-inducing penalty terms. Then penalty terms considered are LASSO-based and the solutions are based on soft-thresholding rule. Because of our intent in biclustering of regression models, we want to discuss the regression approach to solve (8), as given in Lee et al. [4]. For fixed 'u', minimization of (8) with respect to (s, v) is equivalent to minimizing (8) with respect to $\tilde{v} = sv$ if

$$\|Y - u\tilde{v}'\|_F^2 + \lambda_v P_2(\tilde{v}) = \|vec(Y') - (I_n \otimes u)\tilde{v}\|_F^2 + \lambda_v P_2(\tilde{v}) \tag{9}$$

where $\otimes$ denotes the Kronecker product. Note for a given 'u', $I_n \otimes u$ acts as the design matrix and $\tilde{v}$ is the regression coefficient. Note $P_2(\tilde{v}) = \sum_{j=1}^{n} |\tilde{v}_j|$ is the LASSO-penalty.

To solve for 'v', fix $\tilde{u} = su$ and

$$\|Y - \tilde{u}v'\|_F^2 + \lambda_u P_1(\tilde{u}) = \|vec(Y) - (I_m \otimes v)\tilde{u}\|_F^2 + \lambda_u P_1(\tilde{u}) \tag{10}$$

with the LASSO penalty, $P_1(\tilde{u}) = \sum_{i=1}^{m} |\tilde{u}_i|$. Interestingly this regression approach is inherent in the solution of the classical Eckart-Young theorem, that is applied to arrive at the components of singular value decomposition.

4 Parsimonious Regression Models

The main interest in biclustering is to discover some desirable unit-variable association. In the analysis of microarray data, the goal is to identify biologically relevant genes that are significantly expressed for certain cancer types. Initially

the biclustering is used as an unsupervised learning tool; the measurements, the expression levels are for thousands of genes, 'm' over a small number of subjects, 'n'. The information on cancer types is used a posterior to interpret and evaluate the performance of SSVD (see Lee et al. [4]). Visually the low-rank approximations can reveal 'checkerboard' structure resulting from gene and subjects grouping. This feature can be appreciated in many different settings and scientific areas as well.

It is well understood that supervised learning tools fare better as they use the additional information which would also support validating the results. Thus in the context of microarray data, the use of information on the presence or absence of cancer will lead to better clustering; if so, where and how do we look for 'checkerboard' structure? In the regression context the covariances between the response and the predictor variables play an important role and thus the form of regression coefficient matrix matters. We suggest two approaches. Stack up the 'c' coefficients in a matrix and use the SVD on it; we have the $m \times n$ matrix,

$$C \equiv (c_1, \ldots, c_m)' = U\Lambda V' = A \cdot B \tag{11}$$

with rows of U refer to 'm' subjects and the rows of V' refer to the combinations of the predictors, the 'X' variables in the model. When the design matrices X_i are identical, the solution to (11) can be related to reduced-rank regression but when they are different, the calculation of $A(m \times r)$ and $B(r \times n)$ is not straightforward. (See (Chapter 7) in Reinsel and Velu [7]). We provide some essential details here.

With the model as stated in (1) and with the condition on the stacked coefficient matrix, C as stated in (11), it follows that we want to estimate C under the condition,

$$Rank(C) = r \leq min(m, n) \tag{12}$$

which implies that A and B are full-rank matrices. Notice that the model (1) can be rewritten as

$$Y_i = X_i B' a_i + e_i \tag{13}$$

where each a_i is of dimension $r \times 1$. Also notice that BX_i' provides the most useful linear combinations of the predictors and the distinction among the regression equations in (1) are reflected in the coefficients a_i. The dimension-reductions through reduced rank seemed reasonable because of the 'similarity' of the predictor variable sets (x_i) among the m different regression equations in (1). To move forward, we need to impose some normalization conditions (as in normalization required in SVD on U and V):

$$A'\Gamma A = I_r, B\hat{\Sigma}_{xx}B' = \Lambda_r^2 \tag{14}$$

where $\Gamma = \hat{\Sigma}_{ee}^{-1}$ and $\hat{\Sigma}_{xx} = \frac{1}{mT}\sum_{i=1}^{m} X_i'X_i$. To set up the estimation criterion and the resulting estimates, we closely follow the details given in Reinsel and Velu [7].

Observe that the model (1) can be expressed in the vector form as

$$\mathbf{y} = \bar{X}\mathbf{c} + \mathbf{e} \tag{15}$$

where $\mathbf{y} = (Y_1', \ldots, Y_m')'$, $\bar{X} = Diag(X_1, \ldots, X_m)$ and $\mathbf{c} = vec(C')$ with $\mathbf{e} = (e_1', \ldots, e_m')'$, $Cov(\mathbf{e}) = \Sigma_{ee} \otimes I_T$. The generalized least squares (GLS) estimator is given as

$$\hat{\mathbf{c}} = [\bar{X}'(\Sigma_{ee}^{-1} \otimes I_T)\bar{X}]^{-1}\bar{X}'(\Sigma_{ee}^{-1} \otimes I_T)\mathbf{y} \tag{16}$$

The covariance matrix of $\hat{c}$ that is useful to make inferences on c is given as

$$Cov(\hat{\mathbf{c}}) = [\bar{X}'(\Sigma_{ee}^{-1} \otimes I_T)\bar{X}]^{-1} \tag{17}$$

The error covariance matrix is estimated by stocking the residuals $\hat{e_i} = Y_i - X_i\hat{c_i}$ as $\hat{e} = (\hat{e_1}, \ldots, \hat{e_m})'$ and $\hat{\Sigma}_{ee} = \frac{1}{T}\hat{e}\hat{e}'$.

To obtain the estimates of A and B, we need to resort to iterative procedures as there is no one step solution as in SVD. Denote $\alpha = vec(A')$ and $\beta = vec(B')$ and $\theta = (\alpha', \beta')'$. The criterion to be minimized is

$$S_T(\theta) = \frac{1}{2T} \cdot \mathbf{e}'(\Sigma_{ee}^{-1} \otimes I_T)\mathbf{e} \tag{18}$$

subject to the normalizing constraints in (14). Note that $\mathbf{c} = vec(C') = (A \otimes I_n)vec(B') = (I_m \otimes B')vec(A')$ and so $\mathbf{e} = \mathbf{y} - \bar{X}(A \otimes I_n)\beta = \mathbf{y} - \bar{X}(I_m \otimes B')\alpha$. The first order equations that result from minimizing (18) leads to the following iterative solutions—given α solve for β and vice versa.

$$\begin{aligned} \hat{\alpha} &= [\bar{X}(B)'(\Sigma_{ee} \otimes I_T)\bar{X}(B)]^{-1}\bar{X}(B)'(\Sigma_{ee}^{-1} \otimes I_T)\mathbf{y} \\ \hat{\beta} &= [\bar{X}(A)'(\Sigma_{ee} \otimes I_T)\bar{X}(A)]^{-1}\bar{X}(A)'(\Sigma_{ee}^{-1} \otimes I_T)\mathbf{y} \end{aligned} \tag{19}$$

In the above, $\bar{X}(B) = \bar{X}(I_m \otimes B')$ and $\bar{X}(A) = \bar{X}(A \otimes I_n)$.

If the design matrices X_i are the same, the solution is rather straight-forward. We will comment on this model later.

Some observations are in order: Note because $c_i = B'\alpha_i$, which from the estimation point of view implies that,

$$X_i'Y_i = (X_i'X_i)c_i = (X_i'X_i)B'\alpha_i \tag{20}$$

will lead to some simplifications in the cluster 'θ' estimates if we follow the mixtures model approach taken in Sect. 2. Before formulating the problem as finite mixtures with constraints on the regression coefficient matrix, we want to discuss briefly an approach to introduce sparseness in the estimated A and B matrices.

In the first approach we discuss here in similar to Lee et al. [4] where the spareness structure on A and B matrices are introduced and the simplified structure would be used for identifying both the clusters of units and the clusters of the variables. While sparseness studies in the context of reduced-rank regression is of recent origin, many tend to use norms other than Frobenius. In order to keep the focus and for continuity, we will consider Frobenius norm and in that setting, we discuss the decomposition of C-matrix with spareness constraints. Chen and Huang [1] discuss the simultaneous dimension reduction and variable selection. The penalited regression version imposes group-LASSO type penalty

that assumes that each 'c_k' as a group. Note that the range of 'c_k' over various 'k' is not linear space and has certain manifold structure. The methodology provided in this paper is quite straightforward and is easy to implement.

The optimization methods related to reduced-rank regression exploit the bilinear nature of the rank decomposition in (11). Note the coefficient matrix, 'C' is bilinear in the component matries, 'A' and 'B' because given either one, the 'C' matrix is linear function of the other. This leads to the simplified iterative solutions given in (19). Before we formulate the penalized version of the problem, observe from (15)

$$\mathbf{y} = \bar{X} vec(C') + \mathbf{e} = \bar{X}(A \otimes I)B + \mathbf{e} \tag{21}$$

Using (21), the criterion $S_T(\theta)$ in (18) can be restated in terms of approximating full-rank estimate of 'C' matrix by a matrix of reduced rank.

With (16), (17) and (21), the penalized version of the reduced-rank regression model can be stated as,

$$\underset{A,B}{Min} S_T(\theta) + \lambda_A P_1(A) + \lambda_B P_2(B) \tag{22}$$

similar to what is given in (8) for the SSVD of the data matrix Y for one set of variables. As argued in Chen and Huang [1], we will reduce the problem in (22) to minimizing over 'B' for a given 'A', thus resulting in some simplification. First note that minimizing the criterion $S_T(\theta)$ in (18) under the full rank for 'C' matrix is equivalent to

$$\underset{\theta}{Min} S_T^*(\theta) = \frac{1}{2T}(\hat{\mathbf{c}} - \mathbf{c})'[\bar{X}'(\Sigma_{ee}^{-1} \otimes I_T)\bar{X}](\hat{\mathbf{c}} - \mathbf{c}) \tag{23}$$

and thus for a given 'A', it is the same as,

$$\underset{B}{Min} S_T^*(\theta) + \sum_{i=1}^{n} \lambda_i \left\|\beta^i\right\| \tag{24}$$

where 'β^i' denotes the i^{th} row of vector of 'B'' matrix and λ_i's are penalty factors that are positive. The condition is known as group-LASSO which implies that the i^{th} predictor can be taken out of the regression framework. The formulation in (24) is more direct and relates how even in the extended set-up, the problem simplifies to calculating SVD of appropriately weighed full-rank regression coefficient matrix. Observe that minimizing criterion in (24) can be further simplified to, because $\hat{\mathbf{c}} - \mathbf{c} = \hat{\mathbf{c}} - (A \otimes I)\beta$, for a given A, (24) reduces to,

$$\begin{aligned} \underset{\beta}{Min}[(A'\Sigma_{ee}^{-1} \otimes I)\hat{\mathbf{c}} - \beta]'[\bar{X}'(\Sigma_{ee}^{-1} \otimes I_T)\bar{X}] \\ [(A'\Sigma_{ee}^{-1} \otimes I)\hat{\mathbf{c}} - \beta] + \sum_{i=1}^{n} \lambda_i \left\|\beta^i\right\| \end{aligned} \tag{25}$$

Here β^i is the transpose of B^i and denotes the i-th subpart of the 'β' vector.

Solving (25) requires iterative procedures. Before we describe and apply these methods, we want to observe that certain simplifications that occur when the design matrix, $X_i = X$, in a commonly used multivariate regression model. With $\bar{X} = I_T \otimes X$, the optimization in (25) reduces to

$$\begin{aligned} \underset{\beta}{Min}[(A'\Sigma_{ee}^{-1} \otimes I)\hat{\mathbf{c}} - \beta]'[\Sigma_{ee}^{-1} \otimes XX'] \\ [(A'\Sigma_{ee}^{-1} \otimes I)\hat{\mathbf{c}} - \beta] + \sum_{i=1}^{n} \lambda_i \left\|\beta^i\right\|, \end{aligned} \tag{26}$$

the criterion given in Chen and Huang [1]. The solution is easier to obtain because of the simplified structure of

$$\hat{c} = (\Sigma_{ee}^{-1} \otimes XX')^{-1}(\Sigma_{ee}^{-1} \otimes XY') = I_m \otimes (X'X)^{-1}X'Y \tag{27}$$

Our model although appears to be more complex, but it is solvable by numerical routines.

The subgradient solution we suggest here follows Yuan and Lin [8]. The subgradient equations are:

$$[(A'\Sigma_{ee}^{-1} \otimes I)\hat{\mathbf{c}} - \beta]_l + \lambda_l s_l = 0 \tag{28}$$

where $[\cdot]$ denotes the l^{th} subvector of β, where $l = 1, 2, \cdots, n$.

Here $s_l = \beta^l / \left\|\beta^l\right\|$ if $\beta^l \neq 0$ and s_l is a vector with $\|s_l\|_2 < 1$ otherwise. When $\left\|\beta^l\right\| = 0$, to obtain the subgradient equations, inpute 'β' without the 'l^{th}' variable and go through the same process of computations and the soft-threshold estimator is given as:

$$\hat{\beta}^l = \left(1 - \frac{\lambda_l}{\|s_l\|}\right)_+ s_l \tag{29}$$

In our practice, the subgradient algorithm converges in several iterations (less than five) and is not so computationally costly for big datasets.

There may be other ways to obtain this and the research is underway to explore these methods.

5 Parsimonious Finite Mixtures Biclustering

The parsimonious modeling proposed in (22) where both 'A' and 'B' matrices are shrunk through LASSO type penalty may be appropriate to reduce the number of coefficients. But it is not clear how this can be used for grouping. The finite mixture model has the natural appeal as it can be used systematically to decide the number of clusters as well as the probabilities of each unit belonging to various clusters. Recall that in the decomposition the regression coefficient matrix, 'C', the part 'A' represents the units' side and the part 'B' represents variables' side. While the parsimonious or sparse representation may be appropriate, it is

not clear how the spareness in 'A' can easily lead to clustering of observations. So we begin with model (13), where the 'r' dimensional 'a_i' corresponds to unit 'i'. Assume that there are 'K' groups and for a given 'B', thus given $x_i^* = x_i B'$, we postulate that, a_i's come from one of these groups. Thus from (2),

$$\begin{aligned} f(y_i \mid X_i^*, \theta) &= \sum_{k=1}^{K} f(y_i \mid X_i^*, S_i, \theta) f(S_i = k \mid \theta) \\ &= \sum_{k=1}^{K} \eta_k^* f(y_i, X_i^* a_k, \Sigma_k), \\ \textit{where} \quad \theta &= (\eta_1, \ldots, \eta_K, a_1, \ldots, a_K, \Sigma_1, \ldots, \Sigma_K). \end{aligned} \tag{30}$$

This implies that although the row rank of 'A' is taken to be 'm', based on the distance between the rows, it can be reduced to 'K' independent rows.

The EM algorithm (5) can be applied to estimate θ. This is conditional on 'B' given; thus replace 'X_i' in (5) by X_i^*. Now given 'A', estimate 'B' via the second equation in (19). This process can be iterated back and forth between (5) and the estimation of 'β'.

Sparseness: To introduce further sparseness in 'B', we can follow the logic given following equation (21). For a given 'A', use equation (25) and the subgradient method ((28) and (29)) to solve for 'β'. With this the joint modeling of clustering and the dimension reduction are both achieved.

The appropriate model selection as one can see is going to depend upon a number of parameters: rank of the 'C' matrix, 'r' and the row rank of the 'A' matrix, K and the associated other parameters. We will use the grid search via:

$$\begin{aligned} \text{BIC}(r, K) &= 2\text{loglikelihood} - n \ln T \\ \text{BMV}(r, K) &= \max_{r,K}(\text{volume}(\hat{\Sigma}_{r,K})) \end{aligned} \tag{31}$$

This approach is novel and combines both the clustering via finite mixtures and the dimension reduction via SVD. Its performance and its properties are to be studied in depth.

6 Numerical Illustration

The illustration given here is not meant to draw any serious implications toward economic theory but our goal is eventually relate the methodology to draw some useful inferences on the trading behavior on major stocks. Since the early 2000s, both academics and practitioners have paid more attention to the magnitudes of cross-sectional interactions among stocks. However, the study of commonality in short-horizon returns, order flows, and liquidity is still of interest in the microstructure analysis of equity markets. Hasbrouck and Seppi [3] note that both short-horizon returns and order flows are characterized by common factors. With this, two research questions have emerged.

Liquidity commonality can easily arise even when trading activity runs in different directions for different stocks, because sizable buy or sell motivated trading can strain liquidity. But commonality in returns can arise because of less firm-specific and more market-wide factor. For instance public information flows and correlation in order imbalances across market, that may affect all stocks. Furthermore, commonality in order flows may be influenced by the differential liquidity of individual stocks as well as by other factors such as asymmetric information, idiosyncratic risks, transaction costs and other forms of market imperfections.

If commonality in stocks' order flows account for the covariance structure of short-term returns, how should we characterize relationships involving commonality in both returns and order flows? Microstructure research focuses on how individual asset price adjusts to new idiosyncratic information. If the market is efficient, new information would be disseminated and interpreted immediately by all market participants, thus prices would quickly adjust to a new equilibrium value determined by the content of the information. But in practice, the price adjustment does not seem to be processed at the same speed for all stocks. Therefore, the price discovery and order flow dynamics have more complex relationship when we consider multiple assets at the same time.

We chose the Dow stocks as our sample because first, the rapid pace of trading provides frequently updated prices and allows us to construct some high-frequency trading measures. Of the 30 firms in the index, we excluded Kraft Foods, for which data are not available for some duration. Second, these 29 stocks, considered as the large capitalization stocks, are normally categorized in the same style and traded mainly by institutional traders, that is, we can expect more correlated trading on these stocks such as index arbitrage, dynamic hedging strategies and naive momentum trading.

Our sample covers 252 trading days in 2015. We establish a standard time frame for the data series using 15-min intervals covering 9:30–9:45, 9:45–10:00, ..., 15:45–16:00 for a total of 26 intervals per trading day. The 15-min time resolution represents a compromise between, on the one hand, needing to look at correlations in contemporaneous order flows across stocks at very short intervals and, on the other hand, requiring enough time for feedback effects from prices into subsequent order submissions. Such data smoothing is essential in handling high frequency data.

We define the stock returns as the difference between the log end-of-interval quote midpoint and the log begin-of-interval quote midpoint. We also consider the following eight order flow measures: (1) Total number of trades; (2) Total share volume; (3) Total dollar volume using log price as stock price (4) The square root of the dollar volume defined as the sum of the square root of each trade's dollar volume using regular stock price; (5) Signed trades defined as the difference between buy trades and sell trades; (6) Signed share volume defined as the difference between buying share volume and selling share volume; (7) Signed dollar volume defined as the difference between buying dollar volume and selling dollar volume using log price as stock price; (8) The square root of

the signed dollar volume defined as the difference between the sum of the square root of each buy trade's dollar volume and the same measure for each sell trade using regular stock price. These measures are in the 15-min intervals of trading and as the stocks considered are high capitalization stocks, the time intervals always have active observations. These measures generally reflect two dimensions of order flows (See Fig. 1). Table 1 presents two corresponding eigenvectors and shows that the first eigenvector is dominated by "aggregate measures" and the second one is by "signed measures". Thus the first component represents a 'sum' measure and the second component, a 'contrast' measure. The contrast is between two sides, buy and sell.

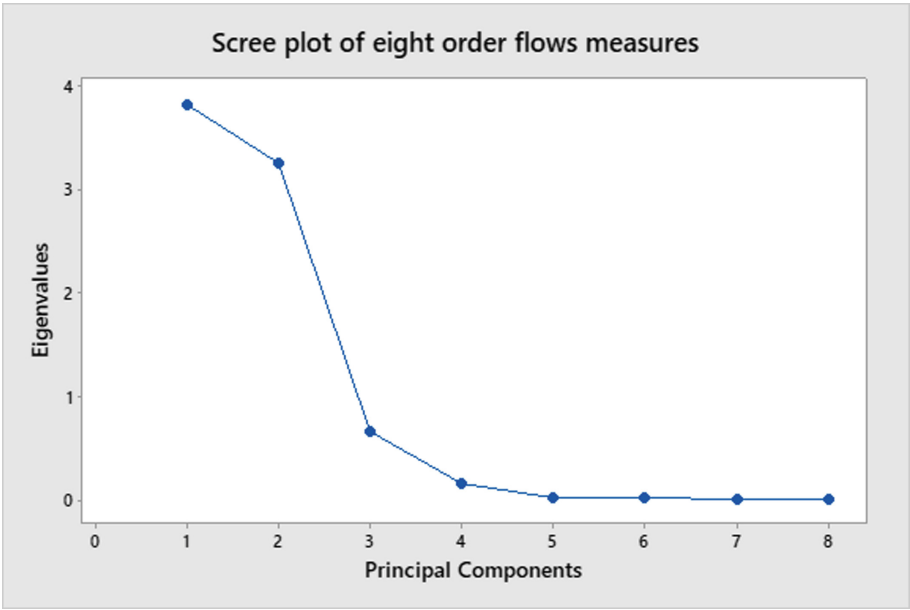

Fig. 1. Scree plot of eight standardized order flows measures

Figure 2 presents the histogram of average returns and averages of eight order flow measures for 29 stocks in the sample. Because these variables have significantly different order of magnitude, we standardize our variables to have unit variance and to remove the time-of-day effects. For a representative variable "z", let $z_{i,d,k}$ denote the observation from firm i on the k-th 15-miniute subperiod of day d. Then the standardized variable becomes $z^*_{i,d,k} = (z_{i,d,k} - \mu_{i,k})/\sigma_{i,k}$, where $\mu_{i,k}$ and $\sigma_{i,k}$ are the mean and standard deviation for firm i and subperiod k, estimated across days.

Table 1. Eigenvectors corresponding to first two largest eigenvalues of eight standardized order flows measures

	Eigenvector 1	Eigenvector 2
Signed trades	−0.099	0.470
Num of trades	0.482	0.117
Signed share volume	−0.127	0.487
Total share volume	0.487	0.113
Signed dollar volume	−0.126	0.487
Total dollar volume	0.486	0.115
Sqrt signed dollar volume	−0.114	0.501
Sqrt total dollar volume	0.490	0.120

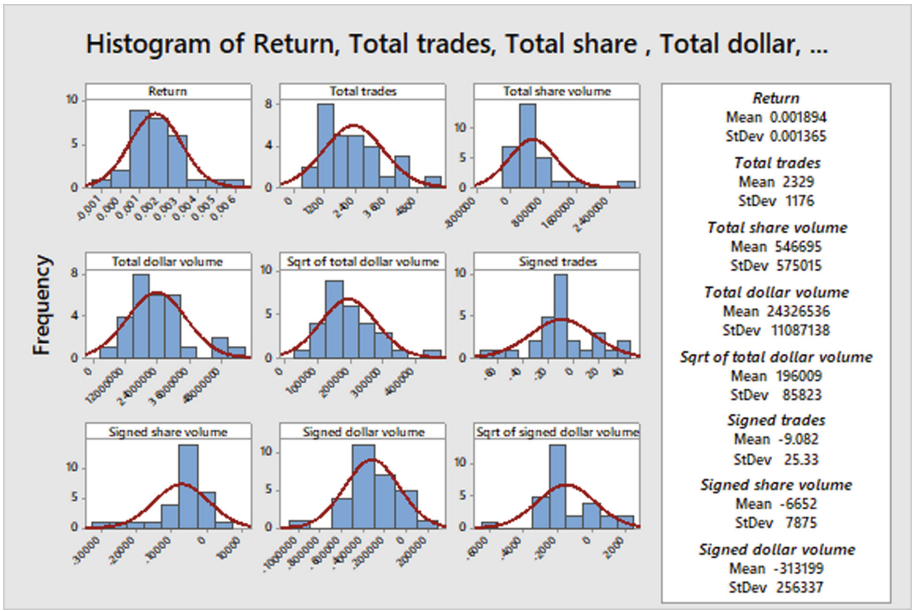

Fig. 2. The histograms of returns mean and 8 order flow measures mean for 29 stocks

We run the following regression for each stock i and set the coefficient equals to 0 if it is insignificant at 0.1 level. The whole 29×8 coefficient matrix is shown in Table 5.

$$r_{i,t} = \sum_{k=1}^{8} c_k x_{k,i,t} + e_i \tag{32}$$

where $r_{i,t}$ is the return for stock i at time t, $x_{k,i,t}$ is the k-th order flow measure for stock i at time t. The coefficient matrix clearly indicates that not all order flow variables are significant. This is an ad-hoc calculation that does not account

for any commonality in the stocks. The methodology developed in this paper provides a more comprehensive framework.

Table 2. Two centroids generated by the K-means clustering ($K = 2$)

Variables	Centorid 1 (1)	Centorid 2 (2)	Difference (2)–(1)
Signed trades	0.203	0.041	−0.162
Num of trades	−0.109	−0.109	0.000
Signed share volume	−0.155	0.811	0.966**
Total share volume	0.020	0.063	0.043
Signed dollar volume	0.159	−0.840	−0.999**
Total dollar volume	−0.029	−0.018	0.011
Sqrt signed dollar volume	0.150	0.454	0.305*
Sqrt total dollar volume	0.097	0.090	−0.007

*, ** denote significant at 5% and 1% level, respectively.

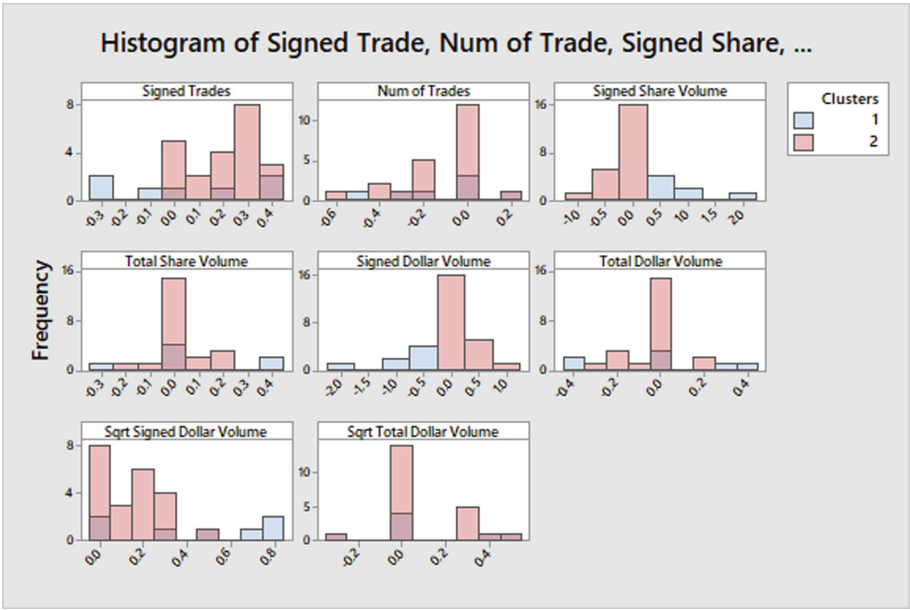

Fig. 3. The histograms of coefficients with 2 clusters

First, financial theory confirms that there are two types traders in the market: noise traders and informed trader, therefore we simply use K-means clustering ($K = 2$) to group the stocks based on the coefficient matrix. Two centroids are shown in Table 2. We observe that the signed share volume and signed dollar

Table 3. Rank two approximation of coefficient matrix using SVD and SSVD algorithm (U matrix)

Company name	SVD		SSVD	
	$U^{(1)}$	$U^{(2)}$	$U^{(1)}$	$U^{(2)}$
Alcoa	−0.139	−0.134	−0.122	0.114
American Express	0.217	−0.211	0.208	0.221
Boeing	0.011	−0.241	0	0.246
Bank of America	−0.212	−0.123	−0.199	0.099
Caterpillar	0.103	−0.050	0.087	0.041
Cisco	−0.275	−0.206	−0.277	0.186
Chevron	−0.010	−0.048	0	0.035
Du Pont	−0.006	−0.082	0	0.070
Disney	0.001	−0.402	0	0.418
General Electric	−0.253	−0.363	−0.255	0.364
Home Depot	0.005	−0.100	0	0.093
Hewlett-Packard	0.010	−0.105	0	0.093
IBM	0.101	−0.054	0.084	0.045
Intel	−0.004	−0.142	0	0.131
Johnson & Johnson	0.352	−0.259	0.354	0.277
JPMorgan Chase	−0.147	−0.383	−0.136	0.379
Coca-Cola	−0.003	−0.141	0	0.134
McDonald	−0.015	−0.083	0	0.069
3M	0.162	−0.239	0.147	0.248
Merck	0.004	−0.022	0	0.011
Microsoft	−0.022	−0.143	0	0.130
Pifzer	0.006	−0.269	0	0.274
Procter & Gamble	−0.021	0.028	0	−0.022
AT& T	−0.716	0.070	−0.738	−0.107
Travelers	0.012	−0.259	0	0.259
United Technologies	0.131	−0.068	0.116	0.061
Verizon	−0.158	0.007	−0.150	0
Wal-Mart	−0.007	−0.016	0	0
Exxon Mobil	−0.008	−0.061	0	0.048

volume have significant difference in these two centroids. It implies that these two variables represent important features for classifying the stocks with different trading behaviors. Figure 3 supports our observation because only the histograms of signed share volume and signed dollar volume have no overlap.

Furthermore, we compute the singular value decomposition (SVD) and the sparse singular value decomposition (SSVD) on the regression coefficients matrix as in (8), (9) and (10). Table 3 shows the results for the U matrix using SVD and SSVD algorithm. When we use K-means clustering(K=2) to cluster these stocks based on U matrix we get from SVD, we obtain the same clustering results as the one using the whole coefficient matrix. It implies that rank-two approximation matrix may have captured most of the information in the structure and the coefficient matrix. From the elements of $U^{(1)}$ from SSVD, we can refer that there may be three groups.

Table 4. Rank two approximation of coefficient matrix using SVD and SSVD algorithm (V matrix)

Variables	SVD		SSVD	
	$V^{(1)}$	$V^{(2)}$	$V^{(1)}$	$V^{(2)}$
Signed trades	0.063	−0.391	0.039	0.383
Num of trades	−0.031	0.516	−0.010	−0.534
Signed share volume	−0.678	0.015	−0.683	0
Total share volume	−0.026	−0.165	−0.006	0.145
Signed dollar volume	0.699	−0.005	0.705	0
Total dollar volume	−0.043	0.275	−0.026	−0.267
Sqrt signed dollar volume	−0.203	−0.448	−0.182	0.426
Sqrt total dollar volume	0.054	−0.526	0.033	0.543

Table 4 presents the 'V' matrix using SVD and SSVD algorithm, we notice that "signed share volume" and "signed dollar volume" dominate all the other variables in both SVD and SSVD cases. Furthermore, it can be observed that more weights are assigned to these two variables when we use the SSVD algorithm.

Table 5. Estimated coefficients of the regressions (32) for all 29 stocks

Company name	Signed trades	Num of trades	Signed share volume	Total share volume	Signed dollar volume	Total share volume	Sqrt signed dollar volume	Sqrt total dollar volume
Alcoa	−0.10*	0	0.32*	0	−0.37*	0	0.70*	0
American Express	0.23*	−0.23*	−0.69*	0	0.70*	0	0.11*	0.28*
Boeing	0.30*	−0.35*	0	0	0	0	0	0.34*
Bank of America	−0.29*	0	0.59*	0	−0.53*	0	0.82*	0
Caterpillar	0.24*	0	−0.30*	0	0.35*	0	0	0
Cisco	0.41*	−0.18*	0.92*	0.40*	−0.93*	−0.40*	0	0
Chevron	0	0	0	0	0	0	0.21*	0
Du Pont	0.19*	0	0	0	0	0	0.20*	0
Disney	0.07	−0.59*	0	0	0	−0.22	0.28*	0.55*
General Electric	0.21*	−0.47*	0.80*	0	−0.86*	0	0.28*	0.55*
Home Depot	0.27*	−0.18	0	0	0	0	0	0
Hewlett-Packard	0.36*	0	−0.07*	0.07	0	0	0.13*	0
IBM	0.26*	0	−0.28	0	0.35*	0	0	0
Intel	0.30*	0	0	0.16	0	−0.18*	0.20*	0
Johnson & Johnson	0	−0.32*	−1.18*	0	1.17*	0	0.33*	0.34*
JPMorgan Chase	0	−0.31*	0.41*	0.35*	−0.44*	−0.45*	0.54*	0.35*
Coca-Cola	0.28*	−0.17	0	0	0	0	0.19*	0
McDonald	0	0	0	0.18	0	0	0.30*	0
3M	0.37*	−0.23*	−0.46*	0	0.52*	−0.20	0	0.28*
Merck	0.34*	0	0	−0.18	0	0.21	0	0
Microsoft	0	0	0	0.12	0	−0.14	0.50*	0
Pifzer	0.29*	−0.35*	0	0	0	0	0.11	0.36*
Procter & Gamble	0	0	0	0	0	0.20*	0.34*	−0.29*
AT& T	−0.31*	0.20*	2.13*	0	−2.21*	0.40*	0.84*	−0.27*
Travelers	0.41*	−0.16	0	0.20	0	−0.32*	0	0.28*
United Technologies	0.32*	0	−0.41*	0	0.41*	0	0	0
Verizon	0.36*	0	0.51	−0.32*	−0.54	0.33*	0	0
Wal-Mart	0.17*	0.20	0	−0.12	0	0	0.20*	0
Exxon Mobil	0.09	0	0	0	0	0	0.20*	0

* coefficients presented are significant at 5% level.

References

1. Chen, L., Huang, J.Z.: Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Am. Stat. **107**(500), 1533–1545 (2012)
2. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, New York (2006). https://doi.org/10.1007/978-0-387-35768-3
3. Hasbrouck, J., Seppi, D.J.: Common factors in prices, order flows, and liquidity. J. Financ. Econ. **59**(3), 383–411 (2001)

4. Lee, M., Shen, H., Huang, J.Z., Marron, J.S.: Biclustering via sparse singular value decomposition. Biometrics **66**(4), 1087–1095 (2010)
5. Qin, L.-X., Self, S.G.: The clustering of regression models method with applications in gene expression data. Biometrics **62**(2), 526–533 (2006)
6. Rao, C.R.: The use and interpretation of principal component analysis in applied research. JSankhyā: Indian J. Stat. Ser. A **26**, 329–358 (1964)
7. Reinsel, G.C., Velu, R.: Multivariate Reduced-Rank Regression: Theory and Applications. Springer, New York (1998). https://doi.org/10.1007/978-1-4757-2853-8
8. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **68**(1), 49–67 (2006)