Sparse-Group Independent Component Analysis with Application to Yield Curves Prediction

Ying Chen[a], Linlin Niu[b], Ray-Bing Chen[c] and Qiang He[d]

[a] *Department of Statistics and Applied Probability, Risk Management Institute,*

National University of Singapore, Singapore

[b] *Wang Yanan Institute for Studies in Economics, Xiamen University, China*

[c] *Department of Statistics, National Cheng Kung University, Taiwan*

[d] *Department of Statistics and Applied Probability, National University of Singapore, Singapore*

July 26, 2018

Abstract

We propose a Sparse-Group Independent Component Analysis (SG-ICA) method to extract independent factors from high dimensional multivariate data. The method provides a unified and flexible framework that automatically identifies the number of factors and simultaneously estimates a sparse loading matrix, enables us to discover important features and offers improved interpretability of the estimators. We establish the consistency and asymptotic normality of the loading matrix estimator, demonstrate its finite sample performance with simulation studies, and illustrate its application using the daily US Overnight Index Swap rates from Oct 2011 to Mar 2015 with 15 maturities ranging from 1 week to 30 years. With higher efficiency of extracting factors, the forecasting performance of the SG-ICA is remarkably better than the popular parametric DNS model in an era of quantitative easing with short-term interest rate

being close to zero.

Keywords: Regularized dimension reduction, Yield curve prediction, Statistically independent factor

1 Introduction

Estimating the dynamic dependence with complex structure is a challenge under the backdrop of big data with large dimensions, which motivates the development of factor identification methods. With a few factors, it makes possible to explain the essential characteristics of large dimensional data in an efficient way. While fundamental models represent factors with exogenous and observable covariates based on theories, statistical methods are data-driven and have attracted much attention without any pre-defined constraints or subjective assumptions. For example, principal component analysis method obtains uncorrelated principal components via covariance eigen-decomposition, see Pearson (1901), Hotelling (1933). Independent component analysis (ICA) method (Hyvärinen et al.; 2001), on the contrary, not only extracts factors through a linear projection but also provides statistical independence among the factors, based on which further investigations can be easily carried out in the univariate space.

The independent components can be estimated using various approaches, including maximising non-Gaussianity (Jones and Sibson; 1987; Cardoso and Souloumiac; 1993; Hyvärinen and Oja; 1997), minimising mutual information (Comon; 1994; Hyvärinen; 1998, 1999a), maximising likelihood (Pham and Garat; 1997; Bell and Sejnowski; 1995; Hyvärinen; 1999b), minimising distance covariance (Matteson and Tsay; 2016). Moreover, the ICA model has been extended to non-linear ICA (Almeida; 2003), kernel ICA (Bach and Jordan; 2003) with nonlinear projection, time-varying coefficient models (Chen et al.; 2014), Bayesian ICA (Winther and Petersen; 2007), nonparametric ICA (Samworth and Yuan; 2012; Chen et al.; 2015) and penalised ICA (Chen et al.; 2017).

Classical ICA models are designed to describe the relation between the observed variables and the latent independent factors via a full-ranked and invertible factor loading matrix. The number of independent components (ICs) equals the dimension of the original multiple variables. Although there is no reduction on dimension, ICA benefits from reducing parameters when modeling the dynamics of the ICs under independence. It is however common in many real-world problems that only a small number of factors are useful in order to explain the stochastic behavior of the original large dimensional data, e.g. the EEG data, (Artoni et al.; 2018). Thus how to perform model selection or regularization on the latent factor identification is an important issue, see Wu et al. (2006).

For a standard linear regression analysis with observed real-valued scalar response and a large number of covariates, many penalty choices are available for variable selection, including the Lasso (Tibshirani; 1996), Ridge (Frank and Friedman; 1993), the smoothly clipped absolute deviation penalty (Fan and Li; 2001) and the adaptive lasso (Zou; 2006). Knight and Fu (2000) proved that the ℓ_1 penalty estimator is consistent under some mild conditions, and the limiting distribution of the Lasso estimator has positive probability mass at the true sparse parameters. Alternatively, there are also penalty functions for group selection. The group Lasso penalty, as an extension of the Lasso penalty, selects a small number of groups, instead of individual factors, by dividing the whole variables into disjoint groups based on the prior information (Bakin; 1999; Antoniadis and Fan; 2001; Cai; 2001; Yuan and Lin; 2006). Obozinski et al. (2011) studied a generalization of the group sparsity, i.e. block-regularization scheme, to identify the active variables in the high-dimensional multivariate linear regression models. However the estimated coefficients are either all zeros or all non-zeros in the same group and thus the group lasso lacks flexibility within group. One can use the Sparse-Group Lasso penalty (Friedman et al.; 2010) that incorporates two-layer regularization, eliminating both insignificant covariates and insignificant coefficients of regression simultaneously, see also Simon et al. (2013), and Chatterjee et al. (2012). A primary challenge in the regularised ICA model is that we do not observe the covariates

and ICs are latent.

In the ICA framework there are two kinds of sparsity either on the source signals or on the mixing matrix (or its inverse). Babaie-Zadeh et al. (2006) proposed a sparse ICA on the signals, see also Khan and Kim (2008) by combining a sparse method and kernel ICA and Bronstein et al. (2005) for implementation on image separation. Abrahamsen and Rigollet (2018) built sparsity on the mixing matrix that is also assumed to be generic. Hyvärinen and Raju (2002) assumed loading matrix is sparse and based on the prior distribution serving as lasso type penalty, the ordinary ICA algorithms are implemented in the estimation. Zhang et al. (2009) adopted the adaptive Lasso penalty for ICA.

In our study we assume sparsity on the loading matrix, i.e. the inverse of the mixing matrix and develop a Sparse-Group Independent Component Analysis (SG-ICA) method, aiming to extract statistically independent factors under sparse group assumption. Compared to the existing work e.g. Hyvärinen and Raju (2002) and Zhang et al. (2009), the proposed method automatically identifies important independent factors without prior distributional knowledge of the number of ICs and simultaneously estimate the sparse loading matrix. The estimation is conducted by maximizing penalized maximum likelihood with the normal inverse Gaussian distributional assumption on the latent ICs. In particular, the Sparse-Group Lasso penalty is imposed to achieve both sparsity in the number of factors and the sparse linear connection. A main contribution of our work is to establish the consistency and asymptotic normality of the group-sparse loading matrix estimator, and demonstrate the finite sample performance of the proposed SG-ICA method with simulation studies. We also implement the SG-ICA method to analyze and forecast the daily US Overnight Index Swap rates from Oct 2011 to Mar 2015 with 15 maturities ranging from 1 week to 30 years. The proposed method enhances the model's interpretability and provides an alternative view to the factors of term structure, with an improved forecast accuracy compared with several alternatives.

The reminder of the paper is structured as follows. Section 2 introduces the SG-

ICA method and describes the estimator of the proposed method. We also present the asymptotic property of the estimators. Section 3 illustrates the usefulness and robustness of the SG-ICA model in practically oriented simulation studies. In Section 3, we implement the SG-ICA to the US overnight index swap rates and show its application in out-of-sample prediction. Section 5 provides concluding remarks. All of the theoretical proofs are contained in the Appendix.

2 SG-ICA

2.1 Classical independent component analysis

Let $\mathbf{X} = (X_1, \cdots, X_p)$ be a p-dimension statistically dependent random variables. The independent component analysis (ICA) model is to factorize the variables into a linear combination of statistically independent random factors $\mathbf{Z} = (Z_1, \cdots, Z_p)$:

$$\mathbf{Z} \quad = \quad B\mathbf{X} \tag{1}$$

where the factor loading matrix B is a $p \times p$ matrix and is assumed to be invertible. In the ICA model, both the factor loadings and the ICs are unknown. Thus the loading matrix and independent components are only identifiable up to scale. For any constant $c \neq 0$, one can obtain another set of loading matrix B/c and ICs $c\mathbf{Z}$ satisfying (1). To avoid the identification problem, the ICs are assumed to have unit variance. The order of ICs on the other hand is not crucial as the ICs are statistically independent and can be marginally analysed without information loss.

2.2 Sparse-group independent component analysis

We now consider regularized independent component analysis with sparse factor loadings and reduced number of factors. Suppose that there are statistically independent random factors $\mathbf{Z} = (Z_1, \cdots, Z_q)$, and the number of factors q is smaller than the

original dimension p:

$$\mathbf{Z} = B_{SG}\mathbf{X} \tag{2}$$

where the factor loading B_{SG} is a $q \times p$ sparse matrix containing a number of zero elements. This is motivated by commonly encountered situations where only a small number of factors are relevant to represent the underlying large dimensional data and the linear dependence between the latent factors and the original data is sparse. Given the observed realizations $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$ with $i = 1, \cdots, n$, the task here is to estimate the sparse loading matrix B_{SG} as well as to obtain the independent factors $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ip})$ with $i = 1, \cdots, n$, without any prior knowledge on the sparsity structure of B_{SG} and the number of ICs q.

In our study, we adopt the maximum likelihood estimation approach. The likelihood is not well-defined without the knowledge of q. Thus we represent the loading matrix in a square form with $p - q$ rows of zeros:

$$B = \begin{bmatrix} B_{SG} \\ N_0 \end{bmatrix} \tag{3}$$

where $N_0 = \mathbf{0}_{(p-q)\times p}$. Unlike in the classical ICA, the square matrix B is singular and is not invertible. It is worth noting that the sequence of factors in ICA can be re-ordered by permutation. Without loss of generality, we suppose that an appropriate permutation has been utilized to move the group-wise zero factor loadings to the bottom of B.

Denote the probability density function of each independent factor to be $f_j(z)$ for $j = 1, \ldots, p$. Then the log-likelihood function is defined as:

$$l(B) = \sum_{i=1}^{n}\sum_{j=1}^{p} \log f_j\left(b_j^\top \mathbf{X}_i\right) + n \log |det(B)|, \tag{4}$$

where $b_j^\top$ denotes the j-th row of B. It is evident that the determinant of the singular matrix B is always zero and thus the likelihood function does not take into account

the magnitude of the loading matrix B_{SG}. We replace N_0 with $N_0^\epsilon = [\mathbf{0} \;\; \epsilon I_{p-q}]$, where $\mathbf{0}$ is a $(p-q) \times q$ zero matrix and ϵ is a pre-specified small constant, and denote the contaminated loading matrix as B^ϵ. A penalty function is introduced to achieve the sparsity and group sparsity of the loading matrix B^ϵ. The penalized log-likelihood function is defined as:

$$\mathbf{P}(B^\epsilon) \;\; = \;\; \sum_{i=1}^{n}\sum_{j=1}^{p} \log f_j(b_j^\top \mathbf{X}_i) + n \log |det(B^\epsilon)| - n\rho_\theta(B^\epsilon) \tag{5}$$

where $b_j^{\epsilon\top}$ denotes the j-th row of B^ϵ, and ρ_θ represents the sparsity penalty with tuning parameter θ. For notational simplicity, we omit the suffix of ϵ in the following description.

2.3 Estimation of loading matrix

Assuming there is only a few number of independent factors and these factors have a sparse connection with the original variables. In other words, only a few rows are active (i.e. not zeros for all elements in the row) in the loading matrix and in each active row some elements are allowed to be zeros. This motivates the adoption of the sparse-group lasso (Friedman et al.; 2010; Simon et al.; 2013) in the penalized optimization of (5). The sparse-group lasso allows both the elementary and group-wise sparsity that involves two penalties given in the form of weighted sum of lasso and group lasso:

$$\rho_{\lambda,\alpha}(B) = \alpha\lambda \sum_{j,k} |b_{jk}| + (1-\alpha)\lambda \sum_{j}^{p} \sqrt{\sum_{k} b_{jk}^2}, \tag{6}$$

where b_{jk} denotes the (j,k)-th element of the sparse loading matrix B. In other words, the sparsity penalty depends on two tuning parameters $\theta = (\lambda, \alpha)$. The parameter λ directly controls the sparsity of elements in the loading matrix. The tuning parameter α adjusts the weights assigned between elementary lasso and group lasso. A large value of λ leads to highly sparse structures, and the importance of elementary sparsity further increases against group sparsity when α enlarges. The performance of regular-

ization depends on the choice of tuning parameters to balance the trade-off between estimation accuracy and loading sparsity. In the penalized regression literature, several selection criteria have been widely applied, including cross validation, generalized cross validation, Mallows' Cp, AIC and BIC (James et al.; 2014). In the framework of SG-ICA where both the loading matrix and ICs are latent, our proposed approach is to choose the tuning parameters using cross-validation.

The density of independent factor is unknown. Motivated by the observations that financial and economic factors are often non-Gaussian distributed with asymmetry and extreme values (Jondeau et al.; 2007), we use the normal inverse Gaussian (NIG) distribution (Barndorff-Nielsen; 1997) for its desirable probabilistic features. With four distribution parameters, the NIG distribution is flexible to mark data characteristics from the central locations to the tails behaviors in a variety of situations. In our study, each independent factor is assumed to be NIG distributed with individual distributional parameters. The density is of the form:

$$f_{\text{NIG}}(z_j) = \frac{\phi_j \delta_j}{\pi} \frac{K_1\Big\{\phi_j \sqrt{\delta_j^2 + (z_j - \mu_j)^2}\Big\}}{\sqrt{\delta_j^2 + (z_j - \mu_j)^2}} \exp\{\delta_j \sqrt{\phi_j^2 - \beta_j^2} + \beta_j (z_j - \mu_j)\},$$

where μ_j, δ_j, β_j and ϕ_j are NIG parameters for $j = 1, \cdots, p$. $K_1(\cdot)$ is the modified Bessel function of the third type. The distributional parameters fulfil the conditions $\mu_j \in \mathbb{R}$, $\delta_j > 0$, and $|\beta_j| \leq \phi_j$. Moreover, all of the independent factors are assumed to have unit variance, avoiding identification ambiguity.

With the NIG distributional assumption, the sparse-group lasso penalized log-

likelihood can be expressed as:

$$\begin{aligned}\mathbf{P}(B) &= \sum_{i}^{n}\sum_{j}^{p}\log f_i(b_j^T\mathbf{X}_i) + n\log|det(B)| - n\rho_{\lambda,\alpha}(B) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{\log\frac{\phi_j\delta_j}{\pi}\frac{K_1\left(\phi_j\sqrt{\delta_j^2+(b_j^\top\mathbf{X}_i-\mu_j)^2}\right)}{\sqrt{\delta_j^2+(b_j^\top\mathbf{X}_i-\mu_j)^2}} + \delta_j\sqrt{\phi_j^2-\beta_j^2} + \beta_j(b_j^\top\mathbf{X}_i-\mu_j)\right\} \\ &\quad +n\log|det(B)| - \alpha\lambda n\sum_{j,k}|b_{jk}| - (1-\alpha)\lambda n\sum_{j}^{p}\sqrt{\sum_{k}^{p}b_{jk}^2} \qquad (7)\end{aligned}$$

The gradient of the penalized log-likelihood function is:

$$\frac{\partial\mathbf{P}}{\partial B} = \sum_{i=1}^{n}\begin{bmatrix}\frac{f_1'(b_1^\top\mathbf{X}_i)}{f_1(b_1^\top\mathbf{X}_i)}\\ \frac{f_2'(b_2^\top\mathbf{X}_i)}{f_2(b_2^\top\mathbf{X}_i)}\\ \cdots\\ \frac{f_p'(b_p^\top\mathbf{X}_i)}{f_p(b_p^\top\mathbf{X}_i)}\end{bmatrix}\mathbf{X}_i^\top + n[B^\top]^{-1} - \Omega$$

where $\Omega_{jk} = \frac{\partial\rho_{\lambda,\alpha}(B)}{\partial b_{jk}}$ and

$$\frac{f_j'(z_j)}{f_j(z_j)} = \beta_j + \phi_j\frac{K_1'(\phi_j\sqrt{\delta_j^2+(z_j-\mu_j)^2})}{K_1(\phi_j\sqrt{\delta_j^2+(z_j-\mu_j)^2})}\frac{z_j-\mu_j}{\sqrt{\delta_j^2+(z_j-\mu_j)^2}} - \frac{z_j-\mu_j}{\delta_j^2+(z_j-\mu_j)^2}.$$

In practice, it is computationally prohibitive to estimate the NIG parameters simultaneously in the penalized estimation of the sparse-group loading matrix B. Therefore the optimization problem is proposed to be solved in two steps, where the maximum is achieved when estimating the loading matrix B given the NIG parameters, while the NIG parameters are learned for a fixed loading matrix, iteratively until the algorithm converges.

Specifically, the algorithm starts with an initial estimator of B_0 from the FastICA method (Hyvärinen; 1999a) using the maximum likelihood estimation under exponential distributional assumption, and then iterates with the following steps:

1. Given the previous estimator of B, optimism the penalized log-likelihood function to obtain the NIG distributional parameters estimator. The EM algorithm is

adopted for the estimation of NIG parameters, see Karlis (2002).

2. Based on the estimated NIG estimator, update the estimator of B by maximizing the penalized log-likelihood function. If any rows of the estimated B are group sparse (all zero), then these rows are kept as zeros and not scaled in step 3.

3. Scale the estimator of B (the not group-sparse rows) and the NIG parameters to have unit variance of each independent factor.

4. Repeat, until convergence achieved.

It is worth mentioning that Hyvärinen (1999b) developed the maximum likelihood estimation approach of independent factor extraction and proved consistency of the estimator under exponential distributional assumption, see also Pham and Garat (1997), and Bell and Sejnowski (1995). The conventional assumption is simple yet unrealistic with only one distributional parameter.

2.4 Asymptotic properties

We derive the asymptotic properties of the SG-ICA estimator under the following three conditions.

C1. The observations $(X_{i1}, \ldots, X_{ip})$ are IID with density $(g_1(X_1, B), \ldots, g_p(X_2, B))$ with respect to some measure μ. The density has a common support and is identifiable. Furthermore, the first logarithmic derivatives of g_a satisfy the equation

$$E\frac{\partial \log g_a(X, B)}{\partial b_{jk}} = 0 \tag{8}$$

for all a, j and k, and

$$E[\frac{\partial \log g_a(X, B)}{\partial b_{j_1 k_1}} \frac{\partial \log g_a(X, B)}{\partial b_{j_2 k_2}}] = E[-\frac{\partial^2 \log g_a(X, B)}{\partial b_{j_1 k_1} \partial b_{j_2 k_2}}] \tag{9}$$

for all a, j_1, j_2, k_1, k_2.

C2. $E[-\Omega_a]$ is positive definite at point B with Ω_a defined as:

$$\Omega_a = \begin{vmatrix} \frac{\partial^2 g_a(B)}{\partial B_{11}\partial B_{11}} & \frac{\partial^2 g_a(B)}{\partial B_{11}\partial B_{12}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{11}\partial B_{1p}} & \frac{\partial^2 g_a(B)}{\partial B_{11}\partial B_{21}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{11}\partial B_{pp}} \\ \frac{\partial^2 g_a(B)}{\partial B_{12}\partial B_{11}} & \frac{\partial^2 g_a(B)}{\partial B_{12}\partial B_{12}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{12}\partial B_{1p}} & \frac{\partial^2 g_a(B)}{\partial B_{12}\partial B_{21}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{12}\partial B_{pp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 g_a(B)}{\partial B_{pp}\partial B_{11}} & \frac{\partial^2 g_a(B)}{\partial B_{pp}\partial B_{12}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{pp}\partial B_{1p}} & \frac{\partial^2 g_a(B)}{\partial B_{pp}\partial B_{21}} & \cdots & \frac{\partial^2 g_a(B)}{\partial B_{pp}\partial B_{pp}} \end{vmatrix}$$

C3. These exists an open subset ω of the parameter space that contains B^{true} such that for almost all x, the density $g_a(X, B)$ admits all third derivatives for all $B \in \omega$. Furthermore, there exist functions $M_{j_1k_1j_2k_2j_3k_3}(x)$ such that the third derivative of the log-density is bounded for all $B \in \omega$:

$$\left|\frac{\partial^3 \log g_a(X, B)}{\partial b_{j_1k_1}\partial b_{j_2k_2}\partial b_{j_3k_3}}\right| \leq M_{j_1k_1j_2k_2j_3k_3}(x). \tag{10}$$

Theorem 1 establishes the estimation consistency by choosing proper penalty parameters.

Theorem 1. Let $(X_{11}, X_{12}, \ldots, X_{1p}), \ldots, (X_{n1}, X_{n2}, \ldots, X_{np})$ be IID measured vector, each with a density $(g_1, g_2, \ldots, g_p)$ that satisfies conditions C1 to C3. If $\max\{\frac{\partial^2 \rho_{\lambda_n,\alpha}(B^{true})}{\partial b_{ij}\partial b_{ij}}\} \to 0$, then there exists a local maximizer $\widehat{B}$ of $\mathbf{P}(B)$ such that $\|\widehat{B} - B^{true}\| = \mathcal{O}_p(n^{-1/2} + a_n)$, where $a_n = \max\{\frac{\partial \rho_{\lambda_n,\alpha}(B^{true})}{\partial b_{ij}}\}$.

In addition, Theorem 2 illustrates that the SG-ICA can identify these zeros elements in the loading matrix with high probability.

Theorem 2. Denote the set $\mathbb{V} = \{(j,k) : b_{jk} \neq 0\}$ and $\mathbb{V}^* = \{(j,k) : b_{jk} = 0\}$. Assume the conditions C1 and C3 are satisfied. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, $\forall B^*$ s.t. $\sqrt{\sum_{(j,k)\in\mathbb{V}}(b^*_{jk} - b_{jk})^2} = O_P(n^{-1/2})$, denote

$$\check{B} = \arg \max_{\sum_{(j,k)\in\mathbb{V}^*} b^2_{jk} \leq C^2/n} \mathbf{P}(B^*).$$

Then with probability tending to 1, $\check{b}_{jk} = 0$ for $(j,k) \in \mathbb{V}^*$

Denote v as the number of elements in $\mathbb{V}$, $\mathbf{b}$ as a vector containing all b_{jk} where $(j,k) \in \mathbb{V}$, $\mathbf{c} = \frac{\partial \rho_{\lambda_n,\alpha}(B^{true})}{\partial \mathbf{b}}$, $K = \frac{\partial^2 \rho_{\lambda_n,\alpha}(B^{true})}{\partial \mathbf{b}^2}$ and $I = -\frac{1}{n} E \frac{\partial^2 \ell(B^{true})}{\partial \mathbf{b}^2}$. Then the asymptotic normality property of the estimation can be found in the following theorem.

Theorem 3. Let $(X_{11}, X_{12}, \ldots, X_{1p}), \ldots, (X_{n1}, X_{n2}, \ldots, X_{np})$ be IID measured vector, each with a density $(g_1, g_2, \ldots, g_p)$ that satisfies condition C1 to C3. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n}-$consistent local maximizer $\widehat{B}$ in Theorem 1 satisfies:

$$\sqrt{n}\{(K - I)(\widehat{\mathbf{b}} - \mathbf{b}^{true}) + \mathbf{c}\} \to N(0, I) \tag{11}$$

The detailed proofs can be found in Appendix.

3 Simulation Studies

We investigate the finite sample performance of the proposed SG-ICA method along with a known data generating process. Our primary interest is on the estimation accuracy under various scenarios. The estimation accuracy refers to the power of detecting active factors and active entries of the loading matrix as well as the deviation of the estimated ICs. We consider a series of simulation studies including dense factor loadings, elementary sparsity and group sparsity with different dimensionality. We first generate $p = 3$ dependent variables under non-sparsity (non-SG) with $q = 3$ independent components, elementary sparsity (ES) with $q = 3$ ICs and sparse group sparsity (SG) with $q = 2$ ICs. Next, we fix the number of ICs $q = 3$ and generate dependent data with dimensionality of $p = 6, 9, 12$ and 15. The simulated data are generated from the model (2) and reflect the real study in the later section. All factor loadings and the ICs' distributional parameters are calibrated from the real-world data of OIS rates from 1^{st} Oct 2011 to 13^{th} Mar 2015 with maturities ranging from one week to 30 years.

Design I is for a moderate number of dependent variables with $p = 3$, with sample

size of $n = 200$. The factor loadings are set as follows:

Non-Sparse (non-SG) scenario:

$$\begin{bmatrix} -0.07 & -0.33 & -0.74 \\ -0.21 & -0.73 & 0.04 \\ -1.63 & 1.15 & -3.11 \end{bmatrix}$$

Elementary sparse (ES) scenario:

$$\begin{bmatrix} \mathbf{0} & -0.33 & -0.72 \\ 1.50 & -0.73 & -0.16 \\ 1.30 & 1.15 & \mathbf{0} \end{bmatrix}$$

Sparse-group (SG) scenario:

$$\begin{bmatrix} \mathbf{0} & -0.33 & -0.72 \\ 1.50 & -0.73 & -0.16 \end{bmatrix}$$

In the scenario the loading matrix is not invertible, and the Moore-Penrose pseudo inverse is used to generate the dependent data. Each scenario is repeated 100 times.

Design II considers the situation where the sparsity increases with the dimension, while the number of ICs is fixed to $q = 3$. The simulation is repeated 100 times with $2,000$ observations each. The factor loadings are as follows:

$p = 6$**:**

$$\begin{bmatrix} & -0.07 & -0.34 & -0.72 & -0.63 & -0.74 \\ 1.50 & -0.21 & -0.73 & -0.16 & -0.33 & 0.04 \\ 1.30 & -1.63 & 1.15 & & & -0.31 \end{bmatrix};$$

$p = 9$**:**

$$\begin{bmatrix} & & & & -0.07 & -0.34 & -0.72 & -0.63 & -0.74 \\ -0.75 & 2.33 & 4.71 & 1.50 & -0.21 & -0.73 & -0.16 & -0.33 & 0.04 \\ 1.47 & -5.25 & 0.46 & 1.30 & -1.63 & 1.15 & & & -0.31 \end{bmatrix};$$

$p = 12$**:**

$$\begin{bmatrix} & & & & & & & -0.07 & -0.34 & -0.72 & -0.63 & -0.74 \\ 1.64 & -3.31 & -1.28 & -0.75 & 2.33 & 4.71 & 1.50 & -0.21 & -0.73 & -0.16 & -0.33 & 0.04 \\ -1.84 & 0.68 & -0.80 & 1.47 & -5.25 & 0.46 & 1.30 & -1.63 & 1.15 & & & -0.31 \end{bmatrix};$$

$p = 15$**:**

$$\begin{bmatrix} & & & & & & & & & & -0.07 & -0.34 & -0.72 & -0.63 & -0.74 \\ & -0.62 & -3.08 & 1.64 & -3.31 & -1.28 & -0.75 & 2.33 & 4.71 & 1.50 & -0.21 & -0.73 & -0.16 & -0.33 & 0.04 \\ -0.46 & 0.94 & 5.51 & -1.84 & 0.68 & -0.80 & 1.47 & -5.25 & 0.46 & 1.30 & -1.63 & 1.15 & & & -0.31 \end{bmatrix}$$

.

We evaluate the estimation accuracy of the SG-ICA method using three measures: the Euclidean distance (ED) indicates the overall estimation accuracy of the loading matrix B_{SG}; the maximum norm (MN) reports the largest elementary bias of the matrix estimator; for the identified independent factors, the root mean squared error (RMSE) is used to show the estimation accuracy:

$$ED = \sum_{jk} \left(b_{jk} - \widehat{b}_{jk} \right)^2 \tag{12}$$

$$MN = \max \left(|b_{jk} - \widehat{b}_{jk}| \right) \tag{13}$$

$$RMSE = \sqrt{\frac{1}{np} \sum_{ij} (z_{ij} - \widehat{z}_{ij})^2} \tag{14}$$

where $\widehat{b}_{jk}$ refers to the estimation of the (j,k)-th element of the matrix B, and $\widehat{z}_{ij}$ is the estimation of the jth component of the independent factor Z_i. For all the three

criteria, the smaller the value, the better is the accuracy. As a comparison, the classical ICA with the FastICA algorithm (Hyvärinen; 1999b) is also considered based on the full rank assumption of B.

Table 1 summaries the simulation results. The proposed SG-ICA is remarkably better than the ICA method, without exception for the three scenarios in Design I. In the elementary and group sparsity scenarios, the proposed SG-ICA provides better results with lower ED, MN and RMSE, accompanied with smaller standard deviations. Even in the non-sparse scenario, the SG-ICA improved accuracy by parsimony in parameter space under sparsity assumption. Moreover, the SG-ICA provides reasonable accuracy in terms of regularization. On average it detects 65% of zeros correctly in the loading matrix for the ES experiment and 62% of zero entries in the loading matrix for the SG experiment. For the latter, the sparse group sparsity is perfectly detected and thus the number of ICs, i.e. $q = 2$, is correctly identified, where the classical ICA fails. When moving from Design I to II, it shows that the performance of the proposed SG-ICA method is stable with respect to increasing dimensionality. For Design II, the SG-ICA method provides even better relative accuracy in all scenarios compared to the classical ICA method. Again, the SG-ICA perfectly identifies the number of ICs and the correct identification on elementary sparsity ranges from 56% to 100%, on average 75%.

4 Application in US Overnight Index Swap Rates

We apply the SG-ICA method to a dataset of the US overnight index swap (OIS) rates obtained from Bloomberg, which contains interest rates at 15 maturities ranging from 1 week to 30 years and covers the period from 1^{st} Oct 2011 to 13^{th} Mar 2015. The OIS rates are the fixed interest rates in exchange for floating interest rates based on the notional swap principal. The referenced floating rates are the US federal funds rates at which depository institutions lend balances to each other overnight. The OIS

Design I: $p = 3, q \leq 3$		ED	MN	RMSE	Group	Entry
non-SG	SG-ICA	0.69(0.27)	0.55(0.24)	0.33(0.20)	-	-
	ICA	1.12(0.55)	2.42(0.65)	0.80(0.11)	-	-
ES	SG-ICA	0.43(0.20)	0.20(0.20)	0.20(0.13)	98%	65%
	ICA	0.53(0.23)	0.31(0.28)	0.27(0.18)	-	-
SG	SG-ICA	0.25(0.13)	0.18(0.12)	0.09(0.03)	100%	62%
	ICA	20.21(6.32)	16.34(5.29)	0.80(0.06)	-	-
Design II: $p \geq 3, q = 3$		ED	MN	RMSE	Group	Entry
$p = 6$	SG-ICA	1.83(0.04)	1.63(0.02)	0.58(0.02)	100%	100%
$q = 3$	ICA	31.90(22.99)	20.40(16.40)	0.93(0.07)	-	-
$p = 9$	SG-ICA	2.28(0.07)	1.20(0.03)	0.12(0.01)	100%	50%
$q = 3$	ICA	15.57(9.48)	11.05(7.42)	0.97(0.02)	-	-
$p = 12$	SG-ICA	6.27(0.20)	4.96(0.35)	0.63(0.03)	100%	56%
$q = 3$	ICA	24.32(8.04)	13.11(6.60)	0.89(0.03)	-	-
$p = 15$	SG-ICA	9.96(0.03)	4.41(0.82)	0.47(0.01)	100%	92%
$q = 3$	ICA	47.73(8.97)	23.75(5.35)	0.73(0.06)	-	-

Table 1: Simulation results for Design I and Design II. ED means Euclidean distance; MN is the maximum norm; RMSE is the root mean squared error; Group refers to the group sparsity defined as the percentage of simulations that recovers the correct number of ICs; and Entry is the elementary sparsity that reports the overall percentage of zero loadings detected correctly. The results of SG-ICA is reported in the first row of each scenario and the results of ICA is reported in the next. For the Non-SG (non-SG) case, the parameters of SG-ICA are $\lambda = 0.1$, $\alpha = 0.4$. For the SG case, the parameter values are $\lambda = 0.06$, $\alpha = 0.3$.

has grown fast in the post 2008 crisis period and gained a major share in the US interest rate derivative market. The OIS rate is close to risk-free rate with negligible credit risk and can be used for the estimation of the yield curve (Sundaresan et al.; 2017). A main goal of our study is to identify and extract statistically independent factors and investigate the impact of factor extraction on the yield curve interpretation and prediction. Figure 1 displays the time evolution of the interest rates. Descriptive statistics of the sample interest rates show that at any maturity the OIS rates are positively skewed and heavily tailed distributed, accompanied with strong linear cross dependence.

4.1 Factors

For the US OIS yield curve data, we extract factors with our SG-ICA approach, as well as the ICA approach. Their forecast performance will be compared. In the mean-

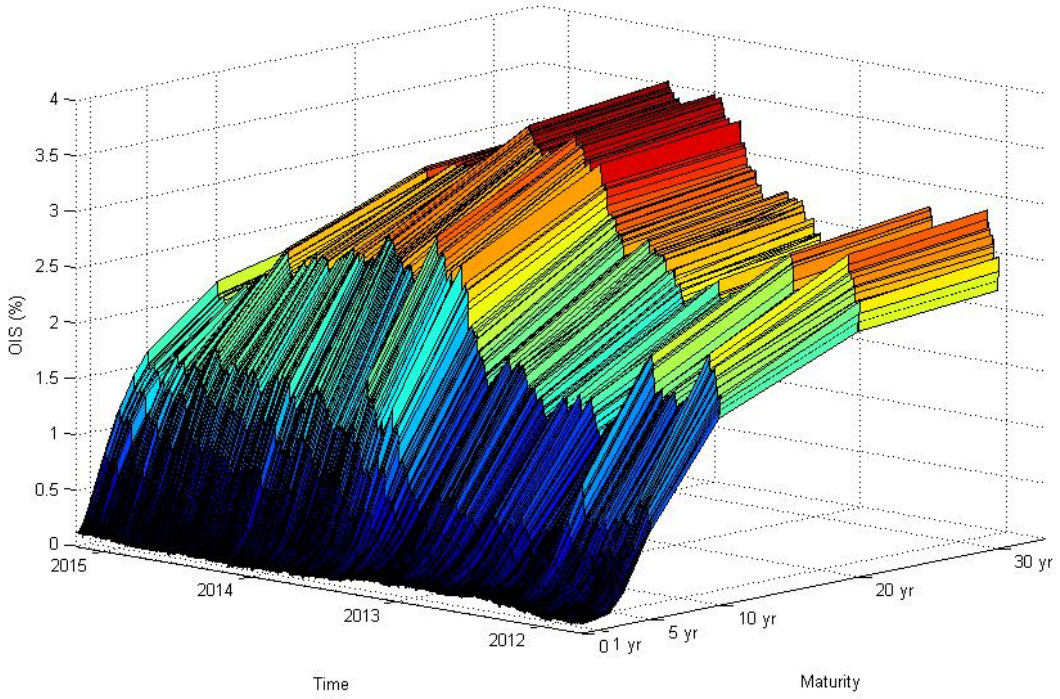

Figure 1: The US OIS rates 2011/10/01-2015/03/13.

while, to provide a comparison with tranditional yield curve models, we also interpolate factors following a parametric approach, i.e., the Nelson-Siegel interpolation. Nelson and Siegel (1987) extract three factors (NS factors hereafter) via a linear projection across yield maturities, which is proved to be very effective in fitting the yield curve of different shapes. The NS factors have also been used in the extended Nelson-Siegel model (Svensson; 1995) and Dynamic Nelson-Siegel (DNS) model (Diebold and Li; 2006). Thanks to its parsimonious setting, simple estimation procedure and superior forecast performance, the DNS model becomes popular among industry experts, central bank researchers and academia (Bank for International Settlements; 2005; Diebold and Rudebusch; 2013). Chen and Niu (2014) demonstrate an outstanding forecast performance of an Adaptive DNS model against a spectrum of alternative yield curve models. Thus, it is natural to compare the factors and forecast performance of the SG-ICA approach and the DNS model, which serves as a good representative of the traditional yield curve models.

The cross-section interpolation in the DNS model can be represented as follows:

$$X_{t,\tau} = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\gamma\tau}}{\gamma\tau}\right) + \beta_{3t}\left(\frac{1-e^{-\gamma\tau}}{\gamma\tau} - e^{-\gamma\tau}\right) + \epsilon_t(\tau), \tag{15}$$

where $X_{t,\tau}$ denotes the yield curve with maturity τ (in months) at time t. The three factors, β_{1t}, β_{2t} and β_{3t}, are named as Level, Slope and Curvature, respectively. Parameter γ controls the exponentially decaying rate of the loadings for the slope and curvature factors, and we follow Diebold and Li (2006) to set $\gamma = 0.0609$ which maximises the curvature loading at a medium maturity of 30 months. We also tried with a different γ estimated from the training sample, but the forecast results are no better than setting it to be 0.0609. So we choose the Diebold and Li (2006) parameterizations.

The SG-ICA method also detects three ICs, choosing the penalty parameters $\lambda = 0.04$ and $\alpha = 0.2$ via cross-validation. Unlike the interpretations of the three NS factors, the ICs are not only estimated in a data-driven manner under independence, but also naturally adapt to features along the dimension of time to maturities. Figure 2

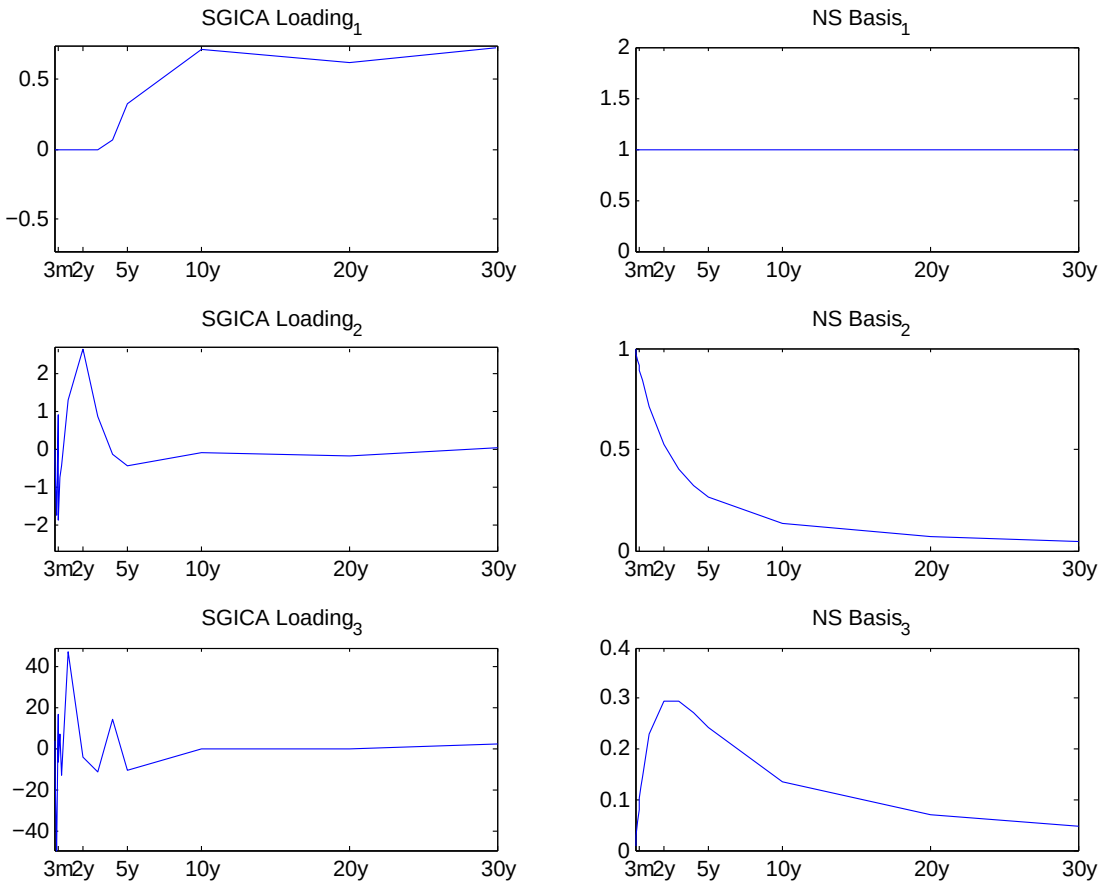

Figure 2: Left panel: SG-ICA loading matrix with $\lambda = 0.04$ and $\alpha = 0.2$. Right panel: NS loadings with $\gamma = 0.0609$

visualises the three loadings of the group sparse matrix B_{SG} against time to maturities from 1 week to 30 years. The NS loadings are displayed on the right panel for comparison. Although the SG-ICA loadings are less smooth compared to those of the NS factors, the figure shows that the first IC contributes to the long term effect of interest rates where the active loadings correspond to $5-$ to $30-$year maturities, the second IC associates with the mid-term effects peaked at the 2-year maturity, and the third one combines both short and mid-term effects manifested on 1- and 4-year maturities.

The time series of the estimated ICs are displayed in Figure 3. Again, the NS factors are displayed on the right panel for comparison. The first row shows that the long term IC and the Level factor of the NS model share similar movement, with a

correlation of 0.98; the second IC shares similar dynamics with Curvature, the third NS factor, with a correlation of 0.87; the third IC is distinctive in its dynamics, and its correlation with any other factor is weak (less than 0.4 in absolute value). Among the NS factors, however, the correlation between Level and Slope are extremely high, at -0.97, indicating that these two factors carry the same information up to a different sign. Thus, in out-of-sample prediction, one of them is redundant conditional on the presence of the other. The redundancy is due to the fact that in recent years since the financial crisis in 2008, the US short term interest rate has been set close to zero for a persistent period as a consequence of the quantitative easing implemented by the US Federal Reserve. Under this backdrop, the Level (close to the long term yield) and the Slope (close to the difference between the short and the long term yields) of the NS model are basically of the same magnitude with opposite signs. By contrast, our ICs not only carry information of the NS factors, but also contain additional information in the third IC, which may add value to forecast.

4.2 Application

Diebold and Li (2006) show that the DNS model does well for out-of-sample forecasts in comparison with various other popular models. We perform an out-of-sample forecast comparison to demonstrate how much the SG-ICA model may improve forecast accuracy when compared with the DNS model.

4.2.1 Forecast procedure

We use the OIS data from 1^{st} Oct 2011 to the end of 2013 as the training sample. We make an out-of-sample forecast in real time and make 1-step ahead forecasts for 1^{st} Jan 2014 to 13^{th} Mar 2015. We move forward one period at a time to re-do the IC estimation and forecast until reaching the end of the sample. Given the feature of statistical independence, it is safe to model and forecast the ICs individually. For fair comparison, we follow Diebold and Li (2006) and consider the autoregressive model of order 1 to describe the dynamics of each IC and update the modelling parameters

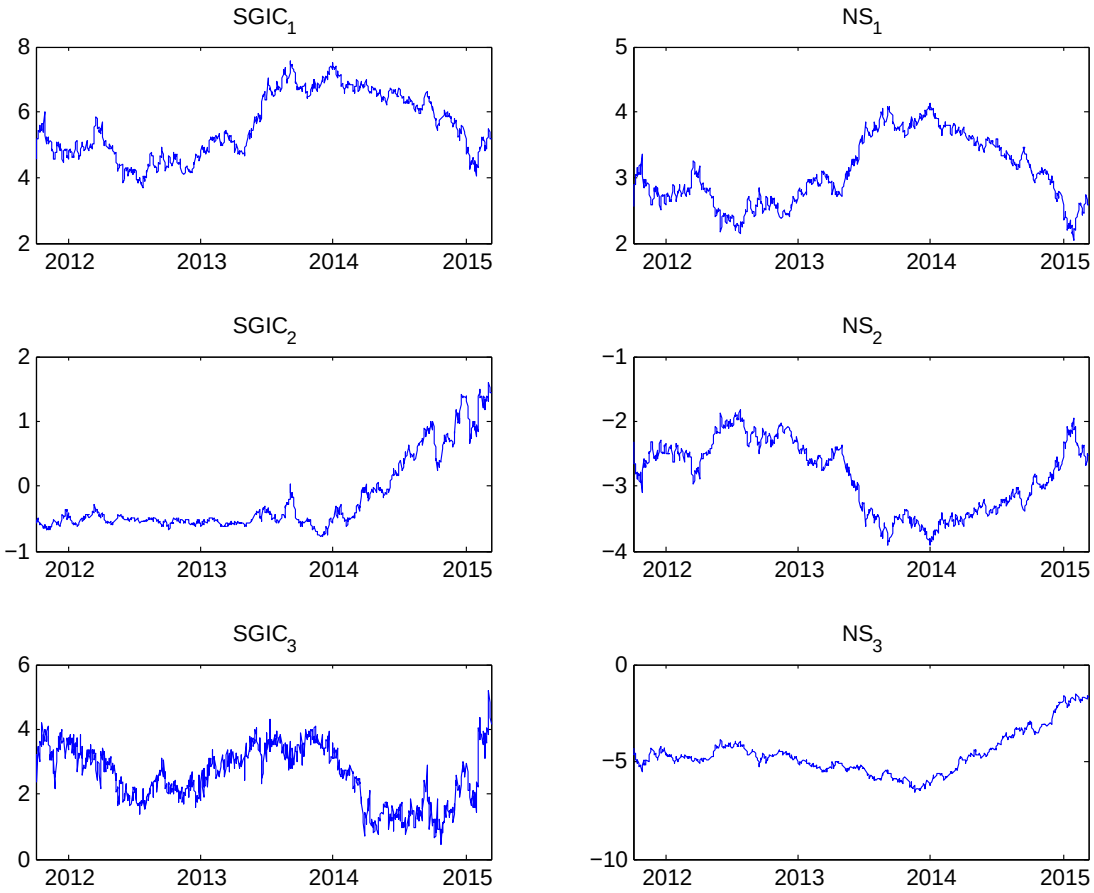

Figure 3: Left panel: ICs in the SG-ICA framework with $\lambda = 0.04$ and $\alpha = 0.2$. Right panel: NS factors with $\gamma = 0.0609$

using a 6−month rolling window:

$$\begin{aligned} \widehat{Z}_{j,t+1} &= \widehat{\varphi}_{0,j,t} + \widehat{\varphi}_{1,j,t} Z_{j,t} \quad j = 1, \ldots, q \\ \widehat{\mathbf{X}}_{t+1} &= \widehat{B}_{SG}^{-1} \widehat{\mathbf{Z}}_{t+1} \quad i = 1, \ldots, p \end{aligned} \tag{16}$$

where $p = 15$ and $q = 3$. Note that the loading matrix $\widehat{B}_{SG}$ is singular and not invertible. Instead of using generalized inverse as $GINV(B_{SG})B_{SG}X \neq X$, the inverse of the loading matrix is estimated in a regression:

$$X_t(\tau) = \phi_{1,\tau} \widehat{Z}_{1,t} + \phi_{2,\tau} \widehat{Z}_{2,t} + \phi_{3,\tau} \widehat{Z}_{3,t} + \epsilon_t$$

where $\phi_{1,\tau}$, $\phi_{2,\tau}$, $\phi_{3,\tau}$ forms the inverse of the loading matrix.

4.2.2 Alternative models for comparison

We choose the ICA model and the DNS model as comparisons. The ICA model is a direct and effective comparison to our SG-ICA model. The DNS model serves as a popular benchmark for yield curve prediction. For the alternative models, we also adopt the 6−month rolling window technique to update the modelling parameters. The rolling estimation reflects the idea of making a trade-off between information efficiency and possible instability in the data generating process.

1) ICA model

We perform one-step-ahead forecast using the ICA model. The forecast procedure of factors and yields is similar to the SG-ICA model, as shown in Equation (16). The essential difference is that the loading matrix is full-ranked and invertible.

2) DNS model

We estimate the DNS model within the NS framework (15) and for each factor we

make the one-step-ahead forecast using an AR(1).

$$\begin{aligned} \widehat{\beta}_{j,t+1} &= \widehat{\varphi}_{0,j,t} + \widehat{\varphi}_{1,j,t}\beta_{j,t} \quad j = 1, \ldots, q \\ \widehat{X}_{t+1,\tau} &= \widehat{\beta}_{1,t+1} + \widehat{\beta}_{2,t+1}\left(\frac{1-e^{-\gamma\tau}}{\gamma\tau}\right) + \widehat{\beta}_{3,t+1}\left(\frac{1-e^{-\gamma\tau}}{\gamma\tau} - e^{-\gamma\tau}\right), \quad i = 1, \ldots, p \end{aligned} \tag{17}$$

The forecast procedure of the NS factors is similar to the SG-ICA model. However the NS factor loadings are fixed with parametric assumption, while the SG-ICA loadings are data-driven.

4.2.3 Measures of forecast comparison

We use the mean absolute forecast error (MAFE) (Ustun and Kasimbeyli; 2012) as the indicator of forecast performance. We calculate the measure for OIS rates in all three models.

4.2.4 Forecast results

Table 2 reports the MAFEs of the interest rates at different maturities. The SG-ICA delivers a reasonable performance for short-term interest rates from 1-week to 6-month, and outperforms the classical ICA without regularization in both mid-term and long-term above 1-year maturity. Compared to the popular DNS model, the data-driven methods provide superior forecast accuracy, with the SG-ICA dominating DNS almost for all maturities. Relating to the information carried in the three factors, the SG-ICA detects more useful information with its parsimonious factors and provides additional prediction power than the NS interpolation, which suffers redundancy in an era of zero policy rate. On average, the forecast accuracy of the SG-ICA (0.022) is remarkably better than the ICA (0.029), and improves upon the DNS forecast precision (0.038) by 42%.

Maturity	SG-ICA	ICA	DNS
1 week	0.010	**0.006**	0.056
1 month	0.009	**0.005**	0.037
2month	0.009	**0.005**	0.016
3month	0.009	**0.005**	0.010
4month	0.009	**0.006**	0.023
5month	0.010	**0.007**	0.035
6month	0.011	**0.008**	0.044
1 year	**0.013**	0.021	0.060
2 year	**0.019**	0.042	0.030
3year	**0.030**	0.050	0.042
4year	**0.039**	0.050	0.051
5year	**0.045**	0.047	0.046
10year	0.038	0.047	**0.037**
20year	**0.034**	0.066	0.046
30year	**0.037**	0.070	0.043
Average	**0.022**	0.029	0.038

Table 2: 1-step ahead out-of-sample forecast accuracy of the SG-ICA and alternatives. The best forecast with the smallest MAFE is marked in bold-face. The second best is underlined.

5 Conclusion

We propose the Sparse-Group Independent Component Analysis (SG-ICA) method to extract independent factors from high dimensional multivariate data. It estimates the loading matrix by maximizing the likelihood function with a sparse group lasso penalty term. Under the assumption that the independent factors are NIG distributed, the proposed method conducts estimation in two steps. With a fixed loading matrix, the NIG parameters are estimated; and then given the NIG distributions, the loading matrix is updated via the sparse group lasso log-likelihood function. The algorithm is iterated until converge. The method provides a unified and flexible framework that automatically identifies the number of factors and simultaneously estimates a sparse loading matrix, enables us to discover important features and offers improved interpretability of the estimators. In addition, our contribution includes the establishment of consistency and asymptotic normality of the loading matrix estimator. We demonstrate the finite sample performance of the SG-ICA method with comprehensive simulation studies, and illustrate its application using the daily US Overnight Index Swap rates from Oct

2011 to Mar 2015 with 15 maturities from 1 week to 30 years. The extracted factors have reasonable interpretations, and simultaneously improves forecast accuracy in mid- and long-term maturities. The forecasting performance of the SG-ICA is remarkably better than the traditional parametric DNS model in an era of quantitative easing with zero policy rates.

While the NIG distribution is assumed in our study, it can be extended for other distributional assumptions customized for the data. Moreover, there may involve correlation in the factors, see e.g. Lee et al. (2011). Thus it is possible to extend the proposed SG-ICA approach by updating the likelihood function (5) with different distributional assumptions.

Acknowledgments

The research of Ying Chen was partly supported by Academic Research Funding R–155–000–178–114 and IDS Funding R–155–000–185–64 at the National University of Singapore and by Natural Science Foundation of China with Grant No. 71528008. The research of Linlin Niu is partly supported by the Natural Science Foundation of China with Grant Nos. 71528008 and 71273007 and Volkswagen Foundation for the project of "QE and Financial (In)stability." The research of Ray-Bing Chen was partially supported by the Ministry of Science and Technology (MOST) of Taiwan and the Mathematics Division of the National Center for Theoretical Sciences in Taiwan.

References

Abrahamsen, N. and Rigollet, P. (2018). Sparse gaussian ica, CoRR **abs/1804.00408**.

Almeida, L. B. (2003). Misep–linear and nonlinear ica based on mutual information, Journal of Machine Learning Research **4**(Dec): 1297–1318.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations, Journal of the American Statistical Association **96**: 939–967.

Artoni, F., Delorme, A. and Makeig, S. (2018). Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition, NeuroImage **175**: 176 – 187.

Babaie-Zadeh, M., Jutten, C. and Mansour, A. (2006). Sparse ica via cluster-wise pca, Neurocomputing **69**(13): 1458 – 1466. Blind Source Separation and Independent Component Analysis.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis, J. Mach. Learn. Res. **3**: 1–48.

Bakin, S. (1999). Adaptive regression and model selection in data mining problems, PhD thesis, School of Mathematical Sciences, Australian National University.

Bank for International Settlements (2005). Zero-coupon Yield Curves: Technical Documentation, BIS papers, Bank for International Settlements, Monetary and Economic Department.

Barndorff-Nielsen, O. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling, Scandinavian Journal of Statistics **24**: 1–13.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution, Neural computation **7**(6): 1129–1159.

Bronstein, A. M., Bronstein, M. M., Zibulevsky, M. and Zeevi, Y. Y. (2005). Sparse ica for blind separation of transmitted and reflected images, International Journal of Imaging Systems and Technology **15**(1): 84–91.

Cai, T. T. (2001). Discussion of regularization of wavelet approximations(by a. antoniadis and j. fan), J. Am. Statist. Ass **96**: 960–962.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non gaussian signals, IEE Proceedings-F **140**: 362–370.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S. and Ganguly, A. (2012). Sparse group lasso: Consistency and climate applications, Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, pp. 47–58.

Chen, R.-B., Chen, Y. and Härdle, W. K. (2014). Tvica-ime varying independent component analysis and its application to financial data, Computational Statistics & Data Analysis **74**: 95–109.

Chen, R.-B., Guo, M., Härdle, W. and Huang, S.-F. (2015). Independent component analysis via copula techniques, Statistics and Computing **25**: 273–288.

Chen, Y., Chen, R. B. and He, Q. (2017). Penalized Independent Factor, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 177–206.

Chen, Y. and Niu, L. (2014). Adaptive dynamic nelson–siegel term structure model with applications, Journal of Econometrics **180**(1): 98–115.

Comon, P. (1994). Independent component analysis, a new concept?, Signal Processing **36**(3): 287 – 314. Higher Order Statistics.

Diebold, F. and Rudebusch, G. (2013). Yield Curve Modeling and Forecasting: The Dynamic Nelson-Siegel Approach, Econometric and Tinbergen Institutes lectures, Princeton University Press.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields, Journal of econometrics **130**(2): 337–364.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association **96**(456): 1348–1360.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, Technometrics **35**(2): 109–135.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso, arXiv preprint arXiv:1001.0736 .

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, Journal of educational psychology **24**(6): 417.

Hyvärinen, A. (1998). Analysis and projection pursuit, Advances in Neural Information Processing Systems **10**: 273.

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis, Neural Networks, IEEE Transactions on **10**(3): 626 –634.

Hyvärinen, A. (1999b). The fixed-point algorithm and maximum likelihood estimation for independent component analysis, Neural Processing Letters **10**(1): 1–5.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). Independent Component Analysis, Wiley-Interscience.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis, Neural computation **9**(7): 1483–1492.

Hyvärinen, A. and Raju, K. (2002). Imposing sparsity on the mixing matrix in independent component analysis, Neurocomputing **49**(1): 151 – 162.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014). An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated.

Jondeau, E., Poon, S. and Rockinger, M. (2007). Financial Modeling Under Non-Gaussian Distributions, Springer Finance, Springer London.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit?, Journal of the Royal Statistical Society. Series A (General) **150**(1): pp. 1–37.

Karlis, D. (2002). An EM type algorithm for maximum likelihood estimation of the normal–inverse gaussian distribution, Statistics & Probability Letters **57**(1): 43–52.

Khan, A. and Kim, I.-T. (2008). Sparse kernel independent component analysis for blind source separation, J. Opt. Soc. Korea **12**(3): 121–125.

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

Lee, S., Shen, H., Truong, Y. K.-N., Lewis, M. and Huang, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging., Journal of the American Statistical Association **106 495**: 1009–1024.

Matteson, D. S. and Tsay, R. S. (2016). Independent component analysis via distance covariance, Journal of the American Statistical Association . Accepted.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves, Journal of business pp. 473–489.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression, The Annals of Statistics pp. 1–47.

Pearson, K. (1901). Principal components analysis, The London, Edinburgh and Dublin Philosophical Magazine and Journal **6**(2): 566.

Pham, D. T. and Garat, P. (1997). Blind separation of mixture of independent sources through a maximum likelihood approach, In Proc. EUSIPCO.

Samworth, R. J. and Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation, The Annals of Statistics **40**(6): 2973–3002.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso, Journal of Computational and Graphical Statistics **22**(2): 231–245.

Sundaresan, S. M., Wang, Z. and Yang, W. (2017). Dynamics of the expectation and risk premium in the ois term structure. Kelley School of Business Research Paper No. 17-41; Columbia Business School Research Paper No. 17-55.

Svensson, L. E. (1995). Estimating forward interest rates with the extended nelson & siegel method, Sveriges Riksbank Quarterly Review **3**(1): 13–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288.

Ustun, O. and Kasimbeyli, R. (2012). Combined forecasts in portfolio optimization: A generalized approach, Computers & OR **39**: 805–819.

Winther, O. and Petersen, K. B. (2007). Bayesian independent component analysis: Variational methods and non-negative decompositions, Digital Signal Processing **17**(5): 858–872.

Wu, E. H., Philip, L. and Li, W. (2006). An independent component ordering and selection procedure based on the MSE criterion, Independent Component Analysis and Blind Signal Separation, Springer, pp. 286–294.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1): 49–67.

Zhang, K., Peng, H., Chan, L. and Hyvärinen, A. (2009). Ica with sparse connections: Revisited, in T. Adali, C. Jutten, J. M. T. Romano and A. K. Barros (eds), Independent Component Analysis and Signal Separation, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 195–202.

Zou, H. (2006). The adaptive lasso and its oracle properties, Journal of the American statistical association **101**(476): 1418–1429.

A Appendix

A.1 Conditions and Proof of Theorem 1

Proof of Theorem 1. For any given $\epsilon > 0$, there exists a large C such that

$$P\{\sup_{\|u\|=C} \mathbf{P}(B^{true} + \alpha_n u) < \mathbf{P}(B^{true})\} \geq 1 - \epsilon, \tag{18}$$

where $\mathbf{P}(B)$ is the penalized likelihood, $\alpha_n = a_n + \frac{1}{\sqrt{n}}$ and u is a p-by-p matrix. Let $D_n(u) = \mathbf{P}(B^{true} + \alpha_n u) - \mathbf{P}(B^{true})$ and

$$I_u(B^{true}) = -E(tr(\nabla_B tr(\nabla \frac{1}{n} l(B^{true})^T u)^T u)) = -E(tr(\nabla_B d_u \frac{1}{n} l(B^{true})^T u)) > 0$$

for any $u \in R^{p*p}$ based on condition C2.

Now we show that $D_n(u) < 0$ by choosing a sufficiently large C, then we can have $\|\widehat{B} - B^{true}\| = O_P(n^{-1/2} + a_n)$, which is equivalent to the proof. Here

$$\begin{aligned}
D(u) &= l(B^{true} + \alpha_n u) - l(B^{true}) - n\{\rho_{\lambda_n}(B^{true} + \alpha_n u) - \rho_{\lambda_n}(B^{true})\} \\
&\leq \alpha_n tr(\nabla l(B)^T u) + \frac{1}{2}\alpha_n^2 tr(\nabla_B d_u l(B)^T u)\{1 + o_P(1)\} \\
&- \{n\alpha_n \sum_{jk} \frac{\partial \rho_{\lambda_n}}{b_{jk}} u_{ik} + \frac{1}{2} n\alpha_n^2 \sum_{j,k,s} \frac{\partial^2 \rho_{\lambda_n}}{\partial b_{jk} \partial b_{js}} u_{jk} u_{js}\{1 + o(1)\}\} \\
&\leq \alpha_n tr(\nabla l(B^{true})^T u) - \frac{1}{2} n\alpha_n^2 I_u(B^{true})\{1 + o_P(1)\} \\
&- \{n\alpha_n \sum_{jk} \frac{\partial \rho_{\lambda_n}}{b_{jk}} u_{ik} + \frac{1}{2} n\alpha_n^2 \sum_{j,k,s} \frac{\partial^2 \rho_{\lambda_n}}{\partial b_{jk} \partial b_{js}} u_{jk} u_{js}\{1 + o(1)\}\}.
\end{aligned} \tag{19}$$

The first inequality is from Taylor expansion and then we substitute $I_u(B)$ into the equation.

Base on condition C1, $n^{-1/2} tr(\nabla l(B^{true})^T u) = O_P(1)$, the first term of Equation (19) is of order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term in $\|u\| = C$.

The last term in Equation (19) is bounded by

$$\sqrt{p}n\alpha_n a_n\|u\| + n\alpha_n^2 \max\{\frac{\partial^2\rho_{\lambda_n}}{\partial b_{jk}\partial b_{js}}\}\|u\|^2. \tag{20}$$

The second term in (20) is also dominated by the second term in (19) as $\max\{\frac{\partial^2\rho_{\lambda_n}(B^{true})}{\partial b_{ij}\partial b_{ij}}\} \to 0$.

Proof is completed.

Proof of Theorem 2. The sufficient condition of Theorem 2 is with probability tending to 1 as $n \to \infty$, $\forall B^*$ such that $\sqrt{\sum_{(j,k)\in\mathbb{V}}(b^*_{jk} - b_{jk})^2} = O_P(n^{-1/2})$, $\epsilon_n = Cn^{-1/2}$ and $(j,k) \in \mathbb{V}^*$

$$\begin{aligned}\frac{\partial \mathbf{P}(B^*)}{\partial b^*_{jk}} &< 0 \quad \text{for} \quad 0 < b^*_{jk} < \epsilon_n;\\ &> 0 \quad \text{for} \quad -\epsilon_n < b^*_{jk} < 0.\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}\frac{\partial \mathbf{P}(B^*)}{\partial b^*_{jk}} &= \frac{\partial l(B^*)}{\partial b^*_{jk}} - n\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}}\\ &= \frac{\partial l(B)}{\partial b^*_{jk}} + \sum_{j_1,k_1}\frac{\partial^2 l(B)}{\partial b^*_{jk}\partial b^*_{j_1k_1}}(b^*_{j_1k_1} - b_{j_1k_1}) + \sum_{j_1k_1}\sum_{j_2k_2}\frac{\partial^3 l(\ddot{B})}{\partial b^*_{jk}\partial b^*_{j_1k_1}\partial b^*_{j_2k_2}}\\ &\quad \times(b^*_{j_1k_1} - b_{j_1k_1})(b^*_{j_2k_2} - b_{j_2k_2}) - n\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}},\end{aligned}$$

where $\ddot{B}$ lies between B^* and B. By standard arguments, we have

$$n^{-1}\frac{\partial l(B)}{\partial b^*_{jk}} = O_P(n^{-1/2})$$

and

$$n^{-1}\frac{\partial^2 l(B)}{\partial b^*_{jk}\partial b^*_{j_1k_1}} = E\{\frac{\partial^2 g(B)}{\partial b^*_{jk}\partial b^*_{j_1k_1}}\} + o_P(1).$$

Under the assumption that $\sqrt{\sum_{(j,k)\in\mathbb{V}}(b^*_{jk}-b_{jk})^2}=O_P(n^{-1/2})$,

$$\frac{\partial \mathbf{P}(B^*)}{\partial b^*_{jk}} = n\lambda_n\{-\lambda_n^{-1}\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}} + O_P(n^{-1/2}/\lambda_n)\} \tag{21}$$

as $n^{-1/2}/\lambda_n \to 0$, the sign of (21) is determined by the sign of $\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}}$, which can be expressed as

$$\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}} = \alpha \times sgn(b^*_{jk}) + (1-\alpha)\frac{b^*_{jk}}{\sqrt{\sum_m b^2_{jm}}}.$$

When $b^*_{jk} > 0$,

$$\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}} = \alpha + (1-\alpha)\frac{b^*_{jk}}{\sqrt{\sum_m b^2_{jm}}} > 0,$$

and when $b^*_{jk} < 0$,

$$\frac{\partial \rho_{\lambda_n}(B^*)}{\partial b^*_{jk}} = -\alpha + (1-\alpha)\frac{b^*_{jk}}{\sqrt{\sum_m b^2_{jm}}} < 0.$$

This completes the proof.

Proof of Theorem 3. By Theorem 2, $\widehat{b}_{jk}=0$ for $(j,k)\in\mathbb{V}^*$. For $(j,k)\in\mathbb{V}$, we have

$$\frac{\partial \mathbf{P}(B)}{\partial b_{jk}}|_{b_{jk}=\widehat{b}_{jk}} = 0 \quad \text{for } (j,k)\in\mathbb{V} \tag{22}$$

Expend the left of (22) and then we have

$$\begin{aligned}
\frac{\partial \mathbf{P}(B)}{\partial b_{jk}}|_{b_{jk}=\widehat{b}_{jk}} &= \frac{\partial \ell(B)}{\partial b_{jk}}|_{b_{jk}=\widehat{b}_{jk}} - n\frac{\partial \rho_{\lambda_n}(\widehat{B})}{\partial b_{jk}} \\
&= \frac{\partial \ell(B^{true})}{\partial b_{jk}} + \sum_{(j_1,k_1)\in\mathbb{V}}\{\frac{\partial^2 \ell(B^{true})}{\partial b_{jk}\partial b_{j_1k_1}} + o_P(1)\}(\widehat{b}_{j_1k_1} - b^{true}_{j_1k_1}) \\
&\quad -n(\frac{\partial \rho_{\lambda_n}(B^{true})}{\partial b_{jk}} + \sum_{(j_1,k_1)\in\mathbb{V}}\{\frac{\partial^2 \rho_{\lambda_n}(B^{true})}{\partial b_{jk}\partial b_{j_1k_1}} + o_P(1)\}(\widehat{b}_{j_1k_1} - b^{true}_{j_1k_1})).
\end{aligned}$$

Thus based on the matrix format, we have

$$\frac{\partial \ell(B^{true})}{\partial \mathbf{b}} + \{\frac{\partial^2 \ell(B^{true})}{\partial \mathbf{b^2}} - nK + o_P(1)\}(\mathbf{b}^{true} - \widehat{\mathbf{b}}) - n\mathbf{c} = \mathbf{0}.$$

Finally by central limit theorem and Slutsky's theorem, we can obtain

$$\sqrt{n}\{(K-I)(\widehat{\mathbf{b}}-\mathbf{b}^{true})+\mathbf{c}\} \to N(0,I).$$

This completes the proof.