Supplement: Technical Details and Additional Numerical Studies for “Supervised Sparse and Functional Principal Component Analysis”

A Tuning Parameter Selection

In this section, we elaborate on the tuning parameter selection procedures for SupSFPC, which are briefly discussed in Section 3.3 of the main paper. For computational efficiency, we embed the selection procedures in each EM iteration, as in Huang et al. (2009) and Allen (2013). Before presenting more technical details, we summarize the comprehensive SupSFPC algorithm in Algorithm 2.

Algorithm 2 EM Algorithm for SupSFPC with Adaptive Tuning Selection

- 1: Initialize model parameters $\theta^{(0)} = (\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \boldsymbol{\Sigma}_{\mathbf{f}}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
 - 2: Repeat until convergence:
 - (a) **E Step:**
 - Get critical conditional expectations (13), (14), and (15);
 - (b) **M Step:**
 - **for** $k = 1 \cdots r$ **do**
 - * Select $\alpha_k^{(t)}$ from (24);
 - * Set $\lambda_k^{(t)}$ to be (25);
 - * Estimate $\mathbf{v}_k^{(t+1)}$ from (20);
 - **end for**
 - Estimate $\sigma_{\mathbf{e}}^{2(t+1)}$ from (18);
 - **for** $k = 1 \cdots r$ **do**
 - * Select $\gamma_k^{(t)}$ from (26);
 - * Estimate $\mathbf{b}_k^{(t+1)}$ from (22);
 - **end for**
 - Estimate $\boldsymbol{\Sigma}_{\mathbf{f}}^{(t+1)}$ from (19);
-

A.1 Select α and λ

The optimization (20) involves two tuning parameters: $\alpha_k^{(t)}$ and $\lambda_k^{(t)}$. They control the smoothness and the sparsity of the k th estimated loading vector $\mathbf{v}_k^{(t+1)}$, respectively. To select the best values for both tuning parameters simultaneously, one may search over a 2-dimensional tuning grid and use cross validation methods (Zou and Hastie, 2005) or information theoretic criteria (Allen, 2013). However, the searching procedure is computationally intensive, especially when we do not have a good knowledge of the range of different tuning parameters and have to search over a large grid. Moreover, we need to repeat the procedure for different PC layers in every EM iteration. The overall computational cost can be huge.

As a remedy, we propose to select $\alpha_k^{(t)}$ and $\lambda_k^{(t)}$ separately. In particular, we omit the sparsity penalty (i.e., set $\lambda_k^{(t)} = 0$) when selecting the smoothness parameter $\alpha_k^{(t)}$, and vice versa. An advantage of this approach is that the optimization (20) reduces to two well-studied problems: a smoothing spline problem (when $\lambda_k^{(t)} = 0$) and a penalized least square problem (when $\alpha_k^{(t)} = 0$). For each respective problem, the other tuning parameter can be selected adaptively using some computationally efficient methods. We drop the subscripts and the superscripts in (20) for simplicity and discuss in more detail below.

When $\lambda = 0$, (20) becomes

$$\min_{\mathbf{v}} \|\boldsymbol{\beta} - \mathbf{v}\|_2^2 + \alpha \mathbf{v}^T \Omega \mathbf{v} \quad (23)$$

where Ω has an expression that is the same as that in smoothing splines (Green and Silverman, 1994). Therefore, (23) is a smoothing spline problem. For a given $\alpha > 0$, the closed-form solution of (23) is $\hat{\mathbf{v}}_\alpha = \mathbf{H}_\alpha \boldsymbol{\beta}$, where $\mathbf{H}_\alpha = (\mathbf{I} + \alpha \Omega)^{-1}$ is a $p \times p$ hat matrix. Leave-one-out cross validation (LOOCV) is commonly used to select the smoothing parameter α in smoothing splines. Given an α , we leave out one entry of $\boldsymbol{\beta}$ at a time, and solve (23) to get a smooth estimate of $\mathbf{v}$; then we calculate the squared difference between the left-out value in $\boldsymbol{\beta}$ and the corresponding interpolated value in $\mathbf{v}$; we repeat the procedure for all entries of $\boldsymbol{\beta}$ and sum up the squared differences as the LOOCV score for this tuning parameter α . In a candidate tuning set, the one that has the smallest

LOOCV score is the optimal tuning parameter.

Solving (23) multiple times for each α can be computationally expensive. However, Green and Silverman (1994) show that the LOOCV score for smoothing spline problems can be obtained analytically by solving the full optimization problem once as

$$\text{LOOCV}(\alpha) = \frac{1}{p} \sum_{j=1}^p \left(\frac{\beta_j - \hat{v}_{\alpha,j}}{1 - h_{\alpha,jj}} \right)^2, \quad (24)$$

where β_j and $\hat{v}_{\alpha,j}$ are the j th entry of $\boldsymbol{\beta}$ and $\hat{\mathbf{v}}_\alpha$, and $h_{\alpha,jj}$ is the j th diagonal entry of $\mathbf{H}_\alpha$. Therefore, LOOCV is an efficient method for tuning parameter selection in smoothing spline. We adopt LOOCV for selecting α in our algorithm. In practice, we can search over a wide range of candidate values at rather low cost.

Given $\alpha = 0$, (20) reduces to a penalized least square problem:

$$\min_{\mathbf{v}} \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{v}\|_2^2 + \lambda \|\mathbf{v}\|_1,$$

which has an explicit solution $\hat{\mathbf{v}}_\lambda = \mathbf{thres}(\boldsymbol{\beta}, \lambda)$. Namely, $\lambda > 0$ is the shrinkage amount imposed on $\boldsymbol{\beta}$. Given $\mathbf{U}$, we know from the definition that the vector $\boldsymbol{\beta} = (\mathbf{E}^T + \mathbf{v}\mathbf{u}^T)\mathbf{u}/\|\mathbf{u}\|_2^2 = \mathbf{v} + \mathbf{E}^T\mathbf{u}/\|\mathbf{u}\|_2^2$, where $\mathbf{E}$ is the measurement error matrix in Model (5) with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2)$. Namely, $\boldsymbol{\beta}$ can be viewed as the true sparse vector $\mathbf{v}$ plus a noise vector with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2/\|\mathbf{u}\|_2^2)$. To accurately estimate the zero entries in $\mathbf{v}$, a proper threshold is the asymptotically tight upper bound of the expectation of infinity norm of the noise vector, which is $\sqrt{2 \log(p) \sigma_{\mathbf{e}}^2 / \|\mathbf{u}\|_2^2}$ (Yang et al., 2013). In practice, since both $\|\mathbf{u}\|_2^2$ and $\sigma_{\mathbf{e}}^2$ are unknown, we substitute them with estimates from the previous EM iteration. In particular, the approximate optimal value for $\lambda_k^{(t)}$ is

$$\lambda_k^{(t)} = \sqrt{2 \log(p) \sigma_{\mathbf{e}}^{2(t)} / c_k^{(t)}}, \quad (25)$$

where $c_k^{(t)} = \mathbb{E}_{\mathbf{U}|\mathbf{X}, \theta^{(t)}}(\mathbf{u}_k^T \mathbf{u}_k)$. Numerical studies indicate this constant works well.

A.2 Select γ

The tuning parameter $\gamma_k^{(t)}$ in (22) is a LASSO sparsity parameter as we formulate and solve (22) as a LASSO problem. Selection of sparsity parameter in a LASSO problem has been well studied in the literature. See, for example, Wang et al. (2009) and Chand (2012). Among data-driven approaches, BIC is a favorable method due to its theoretical merit and fast computation. In particular, since the coordinate descent algorithm can recover the entire solution path efficiently, using BIC to tune LASSO roughly has the same cost as fitting LASSO with a known parameter. Therefore, the BIC procedure for selecting $\gamma_k^{(t)}$ is suitable to be embedded in the EM iteration. For simplicity, we drop the subscripts and the superscripts in the discussion below.

In (22), the BIC score for a given tuning parameter γ is defined as

$$\text{BIC}(\gamma) = n \log(\text{MSE}_\gamma) + \text{df}_\gamma \log(n), \quad (26)$$

where MSE_γ is the mean residual sum of squares, and df_γ is the degree of freedom of the fitted model corresponding to the tuning parameter γ . The degree of freedom of a LASSO fit has been studied in Zou et al. (2007) for a full-column-rank design matrix, and in Tibshirani et al. (2012) for general design matrices. In our case, the design matrix is $\mathbf{Y}$ where columns are potentially linearly dependent. Therefore, we estimate df_γ according to Tibshirani et al. (2012) as

$$\hat{\text{df}}_\gamma = \text{rank}(\mathbf{Y}_{\mathcal{A}(\gamma)}),$$

where $\mathcal{A}(\gamma)$ is a column index set corresponding to nonzero LASSO estimates at γ , and $\mathbf{Y}_{\mathcal{A}(\gamma)}$ is a submatrix of $\mathbf{Y}$ with columns in $\mathcal{A}(\gamma)$. The value that leads to the smallest BIC score is the selected tuning parameter.

B Derivation of (19)

Note that we require Σ_f to be a diagonal matrix, which is denoted by $\Sigma_f = \text{diag}(\sigma_{f,1}^2, \dots, \sigma_{f,r}^2)$. We can rewrite the first term of the object function in (17)

as

$$\begin{aligned}\mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}}[\mathcal{L}(\mathbf{U})] &\propto -\log \det \boldsymbol{\Sigma}_{\mathbf{f}} - \frac{1}{n} \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}} \{ \text{tr}[(\mathbf{U} - \mathbf{Y}\mathbf{B})\boldsymbol{\Sigma}_{\mathbf{f}}^{-1}(\mathbf{U} - \mathbf{Y}\mathbf{B})^T] \} \\ &= -\sum_{k=1}^r \log \sigma_{\mathbf{f},k}^2 - \frac{1}{n} \sum_{k=1}^r \sigma_{\mathbf{f},k}^{-2} \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}} [(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)^T(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)].\end{aligned}$$

Therefore, the maximization over the diagonal entries of the diagonal matrix $\boldsymbol{\Sigma}_{\mathbf{f}}$ can be separated into r univariate problems, each being

$$\max_{\sigma_{\mathbf{f},k}^2} -\log \sigma_{\mathbf{f},k}^2 - \frac{1}{n} \sigma_{\mathbf{f},k}^{-2} \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}} [(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)^T(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)].$$

It then follows that the optimal solution is

$$\widehat{\sigma_{\mathbf{f},k}^2} = \frac{1}{n} \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}} [(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)^T(\mathbf{u}_k - \mathbf{Y}\mathbf{b}_k)].$$

Therefore, $\widehat{\boldsymbol{\Sigma}_{\mathbf{f}}} = \frac{1}{n} \text{diag} \{ \mathbb{E}_{\mathbf{U}|\mathbf{X},\theta^{(t)}} [(\mathbf{U} - \mathbf{Y}\mathbf{B}^{(t+1)})^T(\mathbf{U} - \mathbf{Y}\mathbf{B}^{(t+1)})] \}$.

C Government Bond Yield Data

In this section, we consider the application of SupSFPC to the government bond yield data also studied in Diebold and Li (2006) and Hays et al. (2012). We use the example to illustrate that when auxiliary data are *irrelevant* to the primary data of interest, SupSFPC can adaptively ignore the supervision effect and perform as well as an unsupervised method.

The primary data contain the end-of-month price quotes for U.S. Treasuries, from January 1985 to December 2000 (192 months). For each month, we consider yields on zero coupon bonds of 18 fixed maturities (imputed if missing) of 1.5, 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108, 120 months. Each month is a sample ($n = 192$) and each maturity is a variable ($p = 18$), resulting in a 192×18 primary data matrix $\mathbf{X}$. The 192 raw yield curves of different maturities are shown in Figure 6, with random coloring. For each sample, we also have the auxiliary monthly index information, which may or may not influence the underlying structure of $\mathbf{X}$. In particular, we treat the monthly indices (converted to dummy variables) for the 192 months as supervision data $\mathbf{Y}$.

Figure 6: Raw yield curves of different maturities from January 1975 to December 2000.

Each column of $\mathbf{X}$ and $\mathbf{Y}$ is centered before applying SupSFPC. We set the rank $r = 2$ as the first 2 principal components of the column-centered $\mathbf{X}$ explain over 99% of the total variation. Then we estimate the SupSFPC model parameters from the data. The fitting procedure took less than 1 second to converge (6 EM iterations). The estimated supervision coefficient matrix $\hat{\mathbf{B}}$ is a zero matrix, meaning the auxiliary monthly index data are not relevant to the underlying structure of the yield data. Namely, the yield curves do not present any strong monthly patterns. This is concordant with our observation from the raw data in Figure 6.

We also compare the loading vectors estimated from SupSFPC with those obtained deterministically from the dynamic Nelson-Siegel (DNS) model (Diebold and Li, 2006), which is designed under prior economic theory guidance. Figure 7 shows the comparison results. The first panel shows the mean yield curve from the data versus the first loading from the DNS model, both representing a long-term factor. The deviance indicates that the constant loading of the DNS model may not be adequate to capture the overall yield trend at different maturities. The other two panels show the comparison between the 2nd and 3rd loading vectors between SupSFPC and DNS, respectively. From an economic point of view, the two pre-specified DNS loadings possess the interpretation of medium-term and short-term effects respectively. The two SupSFPC loadings have similar shapes with the respective DNS loadings, meaning that SupSFPC captures similar yield curve patterns as in the DNS model. However, we note that SupSFPC only uses information in the data

without referring to any economic prior knowledge. Namely, SupSFPC is flexible enough to adapt to the dominant features in the data.

Figure 7: Loadings estimated from SupSFPC (solid line) and the pre-specified loadings from the dynamic Nelson-Siegel model (dashed line).

D Emergency Room Visit Data

We now analyze the patient arrival rate data from Armony et al. (2011), that contain hourly number of patients arriving to the emergency room (ER) of the Rambam Hospital, Israel for 417 consecutive days (from September 10th, 2006 to October 31th, 2007). The goal is to understand underlying patient arrival patterns to better allocate human and medical resources. The 417 raw arrival rate curves are shown in the 1st panel of Figure 8.

Other than the hourly arrival rates, we also know the day-of-week index of each day. In particular, the 2nd-8th panels in Figure 8 show arrival curves grouped by the day-of-week index. It can be seen that different days of a week have distinct arrival patterns. Namely, the day-of-week index may be treated as supervision information as it partially drives the underlying structure of the arrival rates.

Each row of the 417×24 primary data matrix contains hourly arrival rates of a day. We apply a square root transformation to the arrival rate data (i.e., $\sqrt{\text{arrival rate} + 1/4}$) to achieve approximate normality (Brown et al., 2005). Then we column center the data matrix and denote it as $\mathbf{X}$. The supervision data matrix $\mathbf{Y}$ contains 417 day-of-week indices (converted to dummy variables and column centered). The rank is set to be 4 based on the scree plot of the singular values of $\mathbf{X}$.

We apply different methods (SupSFPC, SupSVD, one-way SFPC and PCA) to the data.

Figure 8: Raw arrive rate curves of the hospital ER visit data. The first panel shows the overall curves for 417 consecutive days; the other panels show arrival curves on different days of the week respectively.

The fitting procedure of SupSFPC took about 1 minute to converge (166 EM iterations). Figure 9 shows the loading vectors estimated from the methods. By taking into account the auxiliary day-of-week information and allowing regularization, SupSFPC loadings have superior interpretability. The four loadings of SupSFPC capture major variabilities of arrival data from unknown sources after separating the day-of-week effect. They represent large variations of arrival rates at noon, in the evening, overnight, and in the morning, respectively. The day-of-week structure identified by SupSFPC is shown in Figure 10. To get the curves in the figure, we transform $\mathbf{Y}\hat{\mathbf{B}}\hat{\mathbf{V}}^T$ (where $\hat{\mathbf{B}}$ and $\hat{\mathbf{V}}$ are the SupSFPC parameters estimated from the data) back into the original scale by adding the column mean of $\mathbf{X}$ and applying a square transformation. The recovered low-rank structures resemble the (smoothed) average arrival patterns for different weekdays in Figure 8.

Figure 9: The first 4 loading vectors estimated from SupSFPC, SupSVD, SFPC, and PCA.

Figure 10: The day-of-week structure identified by SupSFPC.

References

- Allen, G. I. (2013). Sparse and functional principal components analysis. *arXiv preprint arXiv:1309.2895*.
- Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2011). Patient flow in hospitals: A data-based queueing-science perspective. *New York University*. Available at: <http://www.stern.nyu.edu/om/faculty/armony/>.

- Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50.
- Chand, S. (2012). On tuning parameter selection of lasso-type methods-a monte carlo study. In *Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on*, pages 120–129. IEEE.
- Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.
- Hays, S., Shen, H., Huang, J. Z., et al. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, 6(3):870–894.
- Huang, J. Z., Shen, H., and Buja, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, 104(488).
- Tibshirani, R. J., Taylor, J., et al. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.
- Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.
- Yang, D., Ma, Z., and Buja, A. (2013). A sparse svd method for high-dimensional data. *Journal of Computational and Graphical Statistics*, (Forthcoming).
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.
- Zou, H., Hastie, T., Tibshirani, R., et al. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192.