Supplement to

Scaled Predictor Envelopes and Partial Least Squares Regression

September 25, 2015

A. Marginal scaling

It is common in chemometrics and other application areas to standardize the predictors marginally, so they have common sample variance of 1, prior to application of PLS. To see the consequences of this in terms of model (4), let $\mathbf{\Delta}$ be a diagonal matrix with diagonal elements $\boldsymbol{\gamma}_k^T \mathbf{\Omega} \boldsymbol{\gamma}_k + \boldsymbol{\gamma}_{0k}^T \mathbf{\Omega}_0 \boldsymbol{\gamma}_{0k}$, where $\boldsymbol{\gamma}_k^T$ and $\boldsymbol{\gamma}_{0k}^T$ are the k-th rows of $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$. Then the diagonal matrix of population predictor standard deviations can be represented as $\mathbf{\Delta}^{1/2}\mathbf{\Lambda}$, and model (4) can be re-expressed in terms of the standardized predictors $\mathbf{X}_S = \mathbf{\Delta}^{-1/2}\mathbf{\Lambda}^{-1}\mathbf{X}$ as

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^T \mathbf{\Gamma}^T \mathbf{\Delta}^{1/2} (\mathbf{X}_S - \boldsymbol{\mu}_{\mathbf{X}_S}) + \boldsymbol{\varepsilon}, \qquad (1) \\ \mathbf{\Sigma}_{\mathbf{X}_S} &= \mathbf{\Delta}^{-1/2} \mathbf{\Gamma} \mathbf{\Omega} \mathbf{\Gamma}^T \mathbf{\Delta}^{-1/2} + \mathbf{\Delta}^{-1/2} \mathbf{\Gamma} \mathbf{\Omega}_0 \mathbf{\Gamma}^T \mathbf{\Delta}^{-1/2}. \end{aligned}$$

From this representation we see that marginal scaling does not necessarily mitigate the scaling issue, but rather induces a different rescaling via $\boldsymbol{\Delta}$. In fact, if no scaling is needed from the original scale so $\boldsymbol{\Lambda} = \mathbf{I}_p$, marginal scaling could actually induce a need for rescaling.

B. Derivation of the SPE estimators

The log likelihood function is

$$\begin{aligned} l = \quad & -\frac{n(p+r)}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2}\operatorname{tr}\left[(\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)^T\right] - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\ & -\frac{1}{2}\operatorname{tr}\left[\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}^T\right]. \end{aligned}$$

First we take derivative of l with respect to $\boldsymbol{\mu}_{\mathbf{Y}}^T$, and set it to zero

$$\frac{\partial l}{\partial \boldsymbol{\mu}_Y^T} = \mathbf{1}_n^T\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \overset{\text{set}}{=} 0.$$

Then we get $\bar{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}} = \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})$. Taking the derivative with respect to $\boldsymbol{\mu}_{\mathbf{X}}^T$, and setting it to zero,

$$\frac{\partial l}{\partial \boldsymbol{\mu}_X^T} = \mathbf{1}_n^T(\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{1}_n^T\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} \overset{\text{set}}{=} 0.$$

Substitute the expression of $\bar{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}}$ into the preceding equality, then we obtain $\hat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{X}}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \bar{\mathbf{Y}}$. Up to a constant, the partially maximized log likelihood is then

$$\begin{aligned} l_1 &= -\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2}\operatorname{tr}(\mathbb{X}_c\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbb{X}_c^T) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\ &\quad -\frac{1}{2}\operatorname{tr}\left\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T\right\}. \end{aligned}$$

Now we take derivative of l_1 with respect to $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$,

$$\frac{\partial l_1}{\partial \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}} = -\frac{n}{2}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \overset{\text{set}}{=} 0,$$

to obtain $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = (\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})/n$ and the next the partially maximized log likelihood

$$l_2 = -\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2}\operatorname{tr}(\mathbb{X}_c\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbb{X}_c^T) - \frac{n}{2}\log|(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})|.$$

Take derivative of l_2 with respect to $\boldsymbol{\eta}$ and set it to zero

$$\begin{aligned} \frac{\partial l_2}{\partial \boldsymbol{\eta}} &= -\frac{n}{2}\frac{\partial}{\partial \boldsymbol{\eta}}\log|(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})| \\ &= n\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbb{X}_c^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})\left[(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^T(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})\right]^{-1} \\ &\overset{\text{set}}{=} 0. \end{aligned}$$

Then the estimator of $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{Y}_c = (\boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbf{S}_{\mathbf{X}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbf{S}_{\mathbf{XY}}.$$

As

$$\begin{aligned}
\mathbb{Y}_c - \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \hat{\boldsymbol{\eta}} &= \mathbb{Y}_c - \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{Y}_c \\
&= \left[\mathbf{I}_n - \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \right] \mathbb{Y}_c \\
&= \left[\mathbf{I}_n - \mathbf{P}_{\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}} \right] \mathbb{Y}_c = \mathbf{Q}_{\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}} \mathbb{Y}_c,
\end{aligned}$$

the partially maximized log likelihood becomes

$$\begin{aligned}
l_3 &= -\frac{n}{2} \log |\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2} \operatorname{tr}(\mathbb{X}_c \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \mathbb{X}_c^T) - \frac{n}{2} \log |\mathbb{Y}_c^T \mathbf{Q}_{\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}} \mathbb{Y}_c| \\
&= -\frac{n}{2} \log |\boldsymbol{\Omega}| - \frac{n}{2} \log |\boldsymbol{\Omega}_0| - n \log |\boldsymbol{\Lambda}| - \frac{1}{2} \operatorname{tr}(\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T) \\
&\quad -\frac{1}{2} \operatorname{tr}(\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}_0^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T) - \frac{n}{2} \log |\mathbb{Y}_c^T \mathbf{Q}_{\mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma}} \mathbb{Y}_c|.
\end{aligned}$$

Now we maximize l_3 over $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$,

$$\begin{aligned}
\frac{\partial l_3}{\partial \boldsymbol{\Omega}} &= -\frac{n}{2} \boldsymbol{\Omega}^{-1} + \frac{1}{2} \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Omega}^{-1} \overset{\text{set}}{=} 0, \\
\frac{\partial l_3}{\partial \boldsymbol{\Omega}_0} &= -\frac{n}{2} \boldsymbol{\Omega}_0^{-1} + \frac{1}{2} \boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1} \mathbb{X}_c^T \mathbb{X}_c \boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Omega}_0^{-1} \overset{\text{set}}{=} 0.
\end{aligned}$$

Then the estimators of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are $\widehat{\boldsymbol{\Omega}} = \boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}$ and $\widehat{\boldsymbol{\Omega}}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0$. Since

$$\begin{aligned}
\mathrm{tr}(\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbb{X}_c^T) &= \mathrm{tr}\left[\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbb{X}_c^T\right] \\
&= \mathrm{tr}(n\mathbf{I}_u) = nu \\
\mathrm{tr}(\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}^{-1}\mathbb{X}_c^T) &= \mathrm{tr}\left[\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}^{-1}\mathbb{X}_c^T\right] \\
&= \mathrm{tr}(n\mathbf{I}_{r-u}) = n(r-u),
\end{aligned}$$

the partially maximized log likelihood is

$$\begin{aligned}
l_4 &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0| - \frac{n}{2}\log|\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}}\mathbb{Y}_c| - n\log|\boldsymbol{\Lambda}| \\
&= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}| \\
&\quad -\frac{n}{2}\log|\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}}\mathbb{Y}_c| - n\log|\boldsymbol{\Lambda}| \\
&= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\mathbf{S}_{\mathbf{X}}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}}\mathbb{Y}_c|.
\end{aligned}$$

The last equality is because that $\log|\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}| = \log|\mathbf{S}_{\mathbf{X}}| - 2\log|\boldsymbol{\Lambda}|$.

Let $\mathbf{W} = \mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{Y}$, then

$$\log|\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_c\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}}\mathbb{Y}_c| = \log|\mathbf{S}_{\mathbf{Y}}| + \log|\mathbf{I}_r - \mathbf{S}_{\mathbf{WX}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{XW}}| + \text{constant}.$$

Maximizing l_4 is equivalent to maximizing

$$\begin{aligned} l_5 &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}| \\ &\quad -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\mathbf{I}_r - \mathbf{S}_{\mathbf{WX}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{XW}}| \\ &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\mathbf{S}_{\mathbf{XW}}\mathbf{S}_{\mathbf{WX}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| \\ &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}(\mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{YX}})\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}|. \end{aligned}$$

Therefore, the objective function to minimize is as given in (5).

C. Identifiability

Recall the notation used in the main text: Letting $\mathbf{A} \in \mathbb{R}^{a\times a}$ be a symmetric matrix, we reserve $\mathbf{C}_a \in \mathbb{R}^{a(a+1)/2\times a^2}$ and $\mathbf{E}_a \in \mathbb{R}^{a^2\times a(a+1)2}$ for the “contraction” and “expansion” matrices that connect the vec and vech operators: $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_a\,\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_a\,\mathrm{vech}(\mathbf{A})$. The column vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{q-1})^T$ contains the $q-1$ unique elements of $\boldsymbol{\Lambda}$, so that $\boldsymbol{\lambda}^T = \mathrm{vec}^T(\boldsymbol{\Lambda})\mathbf{L}$, where $\mathbf{L} = (\mathbf{e}_{r_1+1}\otimes\mathbf{e}_{r_1+1}, \cdots, \mathbf{e}_{p-r_q+1}\otimes\mathbf{e}_{p-r_q+1}) \in \mathbb{R}^{p^2\times(q-1)}$ extracts the $q-1$ scaling parameters from $\mathrm{vec}(\boldsymbol{\lambda})$, $\otimes$ denotes Kronecker product, $\mathbf{e}_i \in \mathbb{R}^{p\times 1}$ contains a 1 in the i-th position and 0 elsewhere. Then the constituent parameters in the SPE model (4) are

$$\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\mathbf{Y}}^T,\ \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}), \boldsymbol{\mu}_{\mathbf{X}}^T, \boldsymbol{\lambda}^T,\ \mathrm{vec}^T(\boldsymbol{\eta}),\ \mathrm{vec}^T(\boldsymbol{\Gamma}),\ \mathrm{vech}^T(\boldsymbol{\Omega}),\ \mathrm{vech}^T(\boldsymbol{\Omega}_0)\}^T.$$

Turing to identifiability, if the SPE model (4) has independent but not necessarily normal errors with finite second moments, $\mathbf{S}_{\mathbf{X}} > 0$ and certain technical conditions are met, then it follows from

Shapiro (1986, Proposition 3.1) that $\boldsymbol{\beta}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\phi})$ are identifiable and $\widehat{\boldsymbol{\beta}}_{\text{spe}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\text{spe}}$ are uniquely defined. In the remainder of this section, we connect our context with Shapiro's result. We match our notations with Shapiro's during the discussion.

Our $\boldsymbol{\phi}$ corresponds to Shapiro's $\boldsymbol{\theta}$. The estimable functions in our context are

$$\mathbf{h}(\boldsymbol{\phi}) = \{\boldsymbol{\mu}_{\mathbf{Y}}^T,\, \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}), \boldsymbol{\mu}_{\mathbf{X}}^T,\, \text{vec}^T(\boldsymbol{\beta}),\, \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\},$$

and $\mathbf{h}(\boldsymbol{\phi})$ corresponds to Shapiro's $\boldsymbol{\xi}$. Shapiro's $\hat{\text{x}}$ corresponds to our

$$\tilde{\mathbf{h}} = \{\bar{\mathbf{Y}}^T,\, \text{vech}^T(\mathbf{S}_{\mathbf{Y}|\mathbf{X}}), \bar{\mathbf{X}}^T,\, \text{vec}^T(\widehat{\boldsymbol{\beta}}_{\text{ols}}),\, \text{vech}^T(\mathbf{S}_{\mathbf{X}})\}.$$

Shapiro's discrepancy function F is our log likelihood function, except we omit a constant factor n:

$$\begin{aligned}
F &= l/n \\
&= -(p+r)\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2n}\,\text{tr}\left[(\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)^T\right] - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\
&\quad -\frac{1}{2n}\,\text{tr}\left[\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\left\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^T - (\mathbb{X} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{X}}^T)\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\right\}^T\right] \\
&= -(p+r)\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| - \frac{1}{2}\,\text{tr}\left[\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\{\mathbf{S}_{\mathbf{X}} + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\}\right] - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\
&\quad -\frac{1}{2}\,\text{tr}\left[\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{Y}|\mathbf{X}} + (\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\{\mathbf{S}_{\mathbf{X}} + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\}\right].
\end{aligned}$$

As F is constructed under normal likelihood function, it satisfies conditions 1 - 4 in Section 3 of Shapiro (1986). Shapiro's $\mathbf{V}$ is $\partial^2 F/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T$ evaluated at $(\boldsymbol{\xi}, \boldsymbol{\xi})$. It correspond to the Fisher information matrix $\mathbf{J}^*$ of $\mathbf{h}$ which equals

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} & 0 & -\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} & 0 & 0 \\ 0 & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1})\mathbf{E}_r & 0 & 0 & 0 \\ -\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} & 0 & \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} + \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\beta}^T & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p \end{pmatrix}.$$

Shapiro's $\boldsymbol{\Delta}$ is the gradient matrix $\partial\boldsymbol{\xi}/\partial\boldsymbol{\theta}$ and it is our $\mathbf{H}^* = (\partial\mathbf{h}/\partial\boldsymbol{\phi})$ which equals

$$\begin{pmatrix} \mathbf{I}_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_{r(r+1)/2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I}_p & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1})\mathbf{L} & \mathbf{I}_r \otimes \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Lambda}^{-1} & 0 & 0 \\ 0 & 0 & 0 & 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p)\mathbf{L} & 0 & \mathbf{H}_{56}^* & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}$$

where for notational convenience

$$\mathbf{H}_{56}^* = 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T). \tag{2}$$

As we assume that $\mathbf{S}_{\mathbf{X}} > 0$, $\mathbf{J}^*$ is full rank, $\mathrm{rank}(\mathbf{H}^{*T}\mathbf{J}^*\mathbf{H}^*) = \mathrm{rank}(\mathbf{H}^*)$ and that Shapiro's regularity condition holds. Therefore, all conditions in Shapiro's Proposition 3.1 are satisfied, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are identifiable, and $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ are unique.

D. Proofs

Proof of Proposition 2.1 and Proposition 2.6: Since there is over-parameterization in $\boldsymbol{\Gamma}$, we apply Proposition 4.1 in Shapiro (1986) to compute the asymptotic variance and prove the consistency of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$. We continue to use the notations defined in the proof of Proposition **??**. Shapiro's Proposition 4.1 has the same conditions as his Proposition 3.1, except that it needs an additional condition that $\sqrt{n}(\tilde{\mathbf{h}} - \mathbf{h})$ has to be asymptotically normal. This requires in part that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})$ converge in distribution to a multivariate normal. Recall that $\widehat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbb{X}_c^T\mathbb{X}_c)^{-1}\mathbb{X}_c^T\mathbb{Y}_c$. Since $(\mathbb{X}_c^T\mathbb{X}_c)/n$ converges in probability to $\boldsymbol{\Sigma}_{\mathbf{X}}$, $n(\mathbb{X}_c^T\mathbb{X}_c)^{-1}$ converges in probability to $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$. As $(\mathbf{Y}, \mathbf{X})$ has finite fourth moment, the sequence $\sqrt{n}(\mathbb{X}_c^T\mathbb{Y}_c/n - \boldsymbol{\Sigma}_{\mathbf{XY}})$ converges in distribution. By Slutsky's theorem, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal distribution. It can be shown similarly that asymptotic distribution of $\sqrt{n}(\tilde{\mathbf{h}} - \mathbf{h})$ is multivariate normal. Therefore the conditions of Proposition 4.1 in Shapiro (1986) are all satisfied. Let $\hat{\mathbf{h}}$ be the SPE estimator of $\mathbf{h}$, then $\hat{\mathbf{h}}$ is a consistent estimator of $\mathbf{h}$, and $\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h})$ is asymptotically normally distributed. As $\{\,\text{vec}^T(\boldsymbol{\beta}),\ \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T$ is part of $\mathbf{h}$, $\widehat{\boldsymbol{\beta}}_{\text{spe}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\text{spe}}$ are consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, and $\sqrt{n}[\{\,\text{vec}^T(\widehat{\boldsymbol{\beta}}_{\text{spe}}),\ \text{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\text{spe}})\}^T - \{\,\text{vec}^T(\boldsymbol{\beta}),\ \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T]$ is asymptotically normally distributed. This establishes Proposition 2.6.

Assuming normality, again according to Proposition 4.1 in Shapiro (1986), the asymptotic variance of the SPE estimator of $\mathbf{h}$ has the form $\mathbf{H}^*(\mathbf{H}^{*T}\mathbf{J}^*\mathbf{H}^*)^\dagger\mathbf{H}^{*T}$. As $\mathbf{J}^*$ and $\mathbf{H}^*$ both have block diagonal structure, the asymptotic variance of $\{\,\text{vec}^T(\widehat{\boldsymbol{\beta}}),\ \text{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})\}^T$ is $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$, where

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p \end{pmatrix},$$

and $\mathbf{H}$ has the form

$$\begin{pmatrix} -(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}\otimes\boldsymbol{\Lambda}^{-1})\mathbf{L} & \mathbf{I}_r\otimes\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma} & \boldsymbol{\eta}^T\otimes\boldsymbol{\Lambda}^{-1} & 0 & 0 \\ 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_o\otimes\mathbf{I}_p)\mathbf{L} & 0 & \mathbf{H}^*_{56} & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\otimes\boldsymbol{\Lambda}\boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}_0\otimes\boldsymbol{\Lambda}\boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix},$$

where $\mathbf{H}^*_{56}$ is given at (2). As $\mathbf{J}^{-1}$ is the asymptotic covariance matrix of the OLS estimator of $\{\,\mathrm{vec}^T(\boldsymbol{\beta}),\ \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T$, and $\mathbf{J}^{-1}-\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T=\mathbf{J}^{-1/2}\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}}\mathbf{J}^{-1/2}\geq 0$, the SPE estimators are more efficient than the OLS estimators.

Let $\mathbf{H}=(\mathbf{H}_1,\mathbf{H}_2)$, where $\mathbf{H}_1$ is the first column of $\mathbf{H}$. We write $\mathbf{H}_2=\mathbf{D}\mathbf{H}_o$, where

$$\mathbf{D}=\begin{pmatrix} \mathbf{I}_r\otimes\boldsymbol{\Lambda}^{-1} & 0 \\ 0 & \mathbf{C}_p(\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda})\mathbf{E}_p \end{pmatrix},$$

and

$$\mathbf{H}_o=\begin{pmatrix} \mathbf{I}_r\otimes\boldsymbol{\Gamma} & \boldsymbol{\eta}^T\otimes\mathbf{I}_p & 0 & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega}\otimes\mathbf{I}_p-\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0\otimes\boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix},$$

then $\mathbf{H}_2(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger\mathbf{H}_2^T=\mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^\dagger\mathbf{H}_o^T\mathbf{D}^T$, where $\mathbf{J}_o=\mathbf{D}^T\mathbf{J}\mathbf{D}$. Let

$$\mathbf{T}=\begin{pmatrix} \mathbf{I}_{p-1} & 0 \\ -(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger\mathbf{H}_2^T\mathbf{J}\mathbf{H}_1 & \mathbf{I}_{p(p+1)/2} \end{pmatrix},$$

then $\mathbf{H}\mathbf{T}=(\mathbf{H}_1-\mathbf{P}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1,\mathbf{H}_2)=(\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1,\mathbf{H}_2)$. As $\mathbf{T}$ is full rank, $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T=$

$\mathbf{H}\mathbf{T}(\mathbf{T}^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{T})^\dagger\mathbf{T}^T\mathbf{H}^T$. Now $\mathbf{T}^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{T}$ is a diagonal matrix. This is because

$$\mathbf{H}_2\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})} = \mathbf{H}_2^T\mathbf{J} - \mathbf{H}_2^T\mathbf{J}\mathbf{P}_{\mathbf{H}_2(\mathbf{J})} = \mathbf{H}_2^T\mathbf{J} - \mathbf{H}_2^T\mathbf{J}\mathbf{H}_2(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger\mathbf{H}_2^T\mathbf{J} = \mathbf{H}_2^T\mathbf{J}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{J}^{\frac{1}{2}}\mathbf{H}_2}\mathbf{J}^{\frac{1}{2}} = 0,$$

then

$$\begin{aligned}
\mathbf{T}^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{T} &= \begin{pmatrix} \mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T \\ \mathbf{H}_2^T \end{pmatrix} \mathbf{J} \begin{pmatrix} \mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 & \mathbf{H}_2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 & \mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{H}_2 \\ \mathbf{H}_2^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 & \mathbf{H}_2^T\mathbf{J}\mathbf{H}_2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 & 0 \\ 0 & \mathbf{H}_2^T\mathbf{J}\mathbf{H}_2 \end{pmatrix}.
\end{aligned}$$

Since $(\mathbf{T}^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{T})^\dagger = \operatorname{bdiag}(\left(\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\right)^\dagger, (\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger)$, we have

$$\begin{aligned}
\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T &= (\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1, \mathbf{H}_2)\operatorname{bdiag}(\left(\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\right)^\dagger, (\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger)(\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1, \mathbf{H}_2)^T \\
&= \mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\left(\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\right)^\dagger\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T + \mathbf{H}_2(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^\dagger\mathbf{H}_2^T \\
&= \mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\left(\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\right)^\dagger\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T + \mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^\dagger\mathbf{H}_o^T\mathbf{D}^T \\
&\equiv \mathbf{A} + \mathbf{B}.
\end{aligned}$$

In $\mathbf{B}$, we notice that the upper left $pr \times pr$ block of $\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^\dagger\mathbf{H}_o^T$ is equal to the asymptotic variance of $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_o)$, which does not depend on scaling $\boldsymbol{\Lambda}$. Hence the upper left $pr \times pr$ block of $\mathbf{B}$ is the asymptotic variance of $(\mathbf{I}_p \otimes \boldsymbol{\Lambda}^{-1})\operatorname{vec}(\widehat{\boldsymbol{\beta}}_o)$, which is the cost of estimating $\boldsymbol{\beta}$ when $\boldsymbol{\Lambda}$ is known. In $\mathbf{A}$,

$$\begin{aligned}
\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T &= \mathbf{I} - \mathbf{H}_2(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^{\dagger}\mathbf{H}_2^T\mathbf{J} = \mathbf{I} - \mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\mathbf{D}^T\mathbf{J} \\
&= \mathbf{I} - \mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\mathbf{D}^T\mathbf{J}\mathbf{D}\mathbf{D}^{-1} = \mathbf{I} - \mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\mathbf{J}_o\mathbf{D}^{-1} \\
&= \mathbf{D}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{D}^{-1}, \\
\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})} &= \mathbf{D}^{-T}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}^T\mathbf{D}^T\mathbf{J}\mathbf{D}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{D}^{-1} = \mathbf{D}^{-T}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}^T\mathbf{J}_o\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{D}^{-1},
\end{aligned}$$

$$\begin{aligned}
\mathbf{D}^{-1}\mathbf{H}_1 &= \operatorname{bdiag}(\mathbf{I}_r \otimes \boldsymbol{\Lambda}, \mathbf{C}_p(\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1})\mathbf{E}_p)\begin{pmatrix} -(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1}) \\ 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p) \end{pmatrix}\mathbf{L} \\
&= \begin{pmatrix} -\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_p \\ 2\mathbf{C}_p(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Lambda}^{-1}) \end{pmatrix}\mathbf{L} \\
&= \operatorname{bdiag}(-\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T \otimes \mathbf{I}_p, 2\mathbf{C}_p(\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p))\begin{pmatrix} \boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_p \\ \mathbf{I}_p \otimes \boldsymbol{\Lambda}^{-1} \end{pmatrix}\mathbf{L} \\
&= \operatorname{bdiag}(-\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T \otimes \mathbf{I}_p, 2\mathbf{C}_p(\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p))\begin{pmatrix} (\mathbf{e}_{r0+1} \otimes \mathbf{e}_{r0+1})\lambda_1^{-1} & \dots & (\mathbf{e}_{p-r_{q-1}+1} \otimes \mathbf{e}_{p-r_{q-1}+1})\lambda_{q-1}^{-1} \\ (\mathbf{e}_{r0+1} \otimes \mathbf{e}_{r0+1})\lambda_1^{-1} & \dots & (\mathbf{e}_{p-r_{q-1}+1} \otimes \mathbf{e}_{p-r_{q-1}+1})\lambda_{q-1}^{-1} \end{pmatrix} \\
&= \operatorname{bdiag}(-\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T \otimes \mathbf{I}_p, 2\mathbf{C}_p(\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p))(\mathbf{1}_2 \otimes \mathbf{L})\boldsymbol{\Lambda}_1^{-1} \equiv \mathbf{K}\boldsymbol{\Lambda}_1^{-1},
\end{aligned}$$

where $\boldsymbol{\Lambda}_1^{-1} = \operatorname{diag}\{\lambda_1^{-1}, \dots, \lambda_{q-1}^{-1}\}$, $\mathbf{K} = \operatorname{bdiag}(-\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T \otimes \mathbf{I}_p, 2\mathbf{C}_p(\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p))(\mathbf{1}_2 \otimes \mathbf{L})$, and $\mathbf{1}_2 =$

$(1,1)^T$. Notice that $\mathbf{K}$ does not depend on $\boldsymbol{\Lambda}$. Since $\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 = \boldsymbol{\Lambda}_1^{-1}\mathbf{K}^T\mathbf{Q}_{\mathbf{H}_o(\mathbf{J})}^T\mathbf{J}_o\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}\boldsymbol{\Lambda}_1^{-1}$,

$$\begin{aligned}\mathbf{A} &= \mathbf{D}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}\boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\Lambda}_1^{-1}\mathbf{K}^T\mathbf{Q}_{\mathbf{H}_o(\mathbf{J})}^T\mathbf{J}_o\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}\boldsymbol{\Lambda}_1^{-1})^{\dagger}\boldsymbol{\Lambda}_1^{-1}\mathbf{K}^T\mathbf{Q}_{\mathbf{H}_o(\mathbf{J})}^T\mathbf{D}^T \\ &= \mathbf{D}\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}(\mathbf{K}^T\mathbf{Q}_{\mathbf{H}_o(\mathbf{J})}^T\mathbf{J}_o\mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K})^{\dagger}\mathbf{K}^T\mathbf{Q}_{\mathbf{H}_o(\mathbf{J})}^T\mathbf{D}^T.\end{aligned}$$

Let $\mathbf{G} = \mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}$, then $\mathbf{A} = \mathbf{D}\mathbf{G}(\mathbf{G}^T\mathbf{J}_o\mathbf{G})^{\dagger}\mathbf{G}^T\mathbf{D}^T$, and the asymptotic variance of $\{\,\mathrm{vec}^T(\widehat{\boldsymbol{\beta}}),\ \mathrm{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})\}^T$ has the form $\mathbf{A}+\mathbf{B} = \mathbf{D}\{\mathbf{G}(\mathbf{G}^T\mathbf{J}_o\mathbf{G})^{\dagger}\mathbf{G}^T + \mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\}\mathbf{D}^T$. This completes the proof of Proposition 2.1.

Proof of Corollary 2.2: As $C = \sqrt{\mathrm{tr}(\mathbf{T}_2^{-1/2}\mathbf{T}_1\mathbf{T}_2^{-1/2})}$, we only need to show that $\mathbf{T}_1 = 0$. From the proof of Proposition 2.1, we know that

$$T_1 = (\mathbf{I}_{pr},0)\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\left(\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T\mathbf{J}\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1\right)^{\dagger}\mathbf{H}_1^T\mathbf{Q}_{\mathbf{H}_2(\mathbf{J})}^T(\mathbf{I}_{pr},0)^T,$$

then it is sufficient to show that $(\mathbf{I}_{pr},0)\mathbf{P}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 = (\mathbf{I}_{pr},0)\mathbf{H}_1$. Recall that $\mathbf{H}_2 = \mathbf{D}\mathbf{H}_o$, then $(\mathbf{I}_{pr},0)\mathbf{P}_{\mathbf{H}_2(\mathbf{J})}\mathbf{H}_1 = (\mathbf{I}_{pr},0)\mathbf{H}_2(\mathbf{H}_2^T\mathbf{J}\mathbf{H}_2)^{\dagger}\mathbf{H}_2^T\mathbf{J}\mathbf{H}_1 = (\mathbf{I}_{pr},0)\mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\mathbf{D}\mathbf{J}\mathbf{H}_1$. Notice that $\mathbf{D}$, $\mathbf{J}$ and $\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T$ are all block diagonal matrices. The upper left $pr \times pr$ blocks of $\mathbf{D}$ and $\mathbf{J}$ are $\mathbf{I}_r \otimes \boldsymbol{\Lambda}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}$. According to Cook et al. (2010), the upper left $pr \times pr$ block of $\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T$ is $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_o^{-1}$ when $\boldsymbol{\Sigma}_o = c\mathbf{I}_p$. The rest of the proof follows by straightforward matrix multiplication.

Proof of Proposition 2.3: We continue to use the notations in the proof of Proposition 2.1, the asymptotic variance of $\{\,\mathrm{vec}^T(\widehat{\boldsymbol{\beta}}),\ \mathrm{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})\}^T$ is $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T$, where $\mathbf{J}$ and $\mathbf{H}$ have dimensions $[rp+p(p+1)/2] \times [rp+p(p+1)/2]$ and $[rp+p(p+1)/2] \times [q-1+ru+u^2+p(p+1)/2]$.

We can write $\mathbf{H} = \mathbf{M}\mathbf{N}$, where the $[rp + p(p+1)/2] \times [q - 1 + ru + p(p+1)/2]$ matrix $\mathbf{M}$ has the form

$$\begin{pmatrix} -(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{\Lambda}^{-1})\mathbf{L} & \mathbf{I}_r \otimes \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}_0 & 0 & 0 \\ 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_o \otimes \mathbf{I}_p)\mathbf{L} & 0 & 2\mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0 - \boldsymbol{\Lambda}\boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Lambda}\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix},$$

and the $[q - 1 + ru + p(p+1)/2] \times [q - 1 + ru + u^2 + p(p+1)/2]$ matrix $\mathbf{N}$ equals

$$\begin{pmatrix} \mathbf{I}_{q-1} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_{ru} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}^T & 0 & 0 \\ 0 & 0 & \mathbf{I}_u \otimes \boldsymbol{\Gamma}_0^T & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_u(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}^T) & \mathbf{I}_{u(u+1)/2} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I}_{(p-u)(p-u+1)/2} \end{pmatrix}.$$

As $\mathbf{N}$ has full row rank, the rank of $\mathbf{H}$ is equal to the rank of $\mathbf{M}$, then the asymptotic variance of $\{\, \mathrm{vec}^T(\widehat{\boldsymbol{\beta}}), \mathrm{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})\}^T$ is $\mathbf{M}(\mathbf{M}^T\mathbf{J}\mathbf{M})^{\dagger}\mathbf{M}^T$. When $u > p - (q-1)/r$, $rp + p(p+1)/2 < q - 1 + ru + p(p+1)/2$, $\mathbf{M}$ has more columns than rows. According to Shapiro (1986), the rank of $\mathbf{M}$ is the number of independent parameters in the model, then the rank of $\mathbf{M}$ should be $rp + p(p+1)/2$. We perform a singular value decomposition to $\mathbf{M}$: $\mathbf{M} = \mathbf{L}\mathbf{D}\mathbf{R}$, where $\mathbf{L} \in \mathbb{R}^{[rp+p(p+1)/2]\times[rp+p(p+1)/2]}$ and $\mathbf{R} \in \mathbb{R}^{[q-1+ru+p(p+1)/2]\times[q-1+ru+p(p+1)/2]}$ are orthogonal matrices, $\mathbf{D} = (\mathbf{D}_0,\ \mathbf{0}) \in \mathbb{R}^{[rp+p(p+1)/2]\times[q-1+ru+p(p+1)/2]}$ and $\mathbf{D}_0 \in \mathbb{R}^{[rp+p(p+1)/2]\times[rp+p(p+1)/2]}$ is a diagonal matrix with non-zero diagonal elements. Then

$$\begin{aligned}\mathbf{M}(\mathbf{M}^T\mathbf{J}\mathbf{M})^\dagger\mathbf{M}^T &= \mathbf{L}\mathbf{D}\mathbf{R}(\mathbf{R}^T\mathbf{D}^T\mathbf{L}^T\mathbf{J}\mathbf{L}\mathbf{D}\mathbf{R})^\dagger\mathbf{R}^T\mathbf{D}^T\mathbf{L}^T\\ &= \mathbf{L}(\mathbf{D}_0,\ \mathbf{0})\mathbf{R}[\mathbf{R}^T(\mathbf{D}_0^{-1},\ \mathbf{0})^T\mathbf{L}^T\mathbf{J}^{-1}\mathbf{L}(\mathbf{D}_0^{-1},\ \mathbf{0})\mathbf{R}]\mathbf{R}^T(\mathbf{D}_0,\ \mathbf{0})^T\mathbf{L}^T\\ &= \mathbf{L}(\mathbf{D}_0,\ \mathbf{0})(\mathbf{D}_0^{-1},\ \mathbf{0})^T\mathbf{L}^T\mathbf{J}^{-1}\mathbf{L}(\mathbf{D}_0^{-1},\ \mathbf{0})(\mathbf{D}_0,\ \mathbf{0})^T\mathbf{L}^T\\ &= \mathbf{L}\mathbf{L}^T\mathbf{J}^{-1}\mathbf{L}\mathbf{L}^T = \mathbf{J}^{-1}.\end{aligned}$$

Note that $\mathbf{J}^{-1}$ is the asymptotic covariance matrix of the OLS estimator of $\{\,\mathrm{vec}^T(\boldsymbol{\beta}),\ \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T$, which establishes Proposition 2.3.

E. Relative cost of estimating $\boldsymbol{\Lambda}$

Under the simulation setup for Figure 2, we varied the signal and noise levels to investigate the relative cost C of estimating $\boldsymbol{\Lambda}$, as defined in Section 2.3. We fixed σ_0 at $\sqrt{5}$ and let σ equal to 0.1, 0.2, 0.5, 1, $\sqrt{5}$, 5 and 10. We took different signal levels from multiplying $\boldsymbol{\eta}$ by 0.2, 1 and 5. A plot of the relative cost C is shown in Figure I. From Figure I, we notice that the cost increases

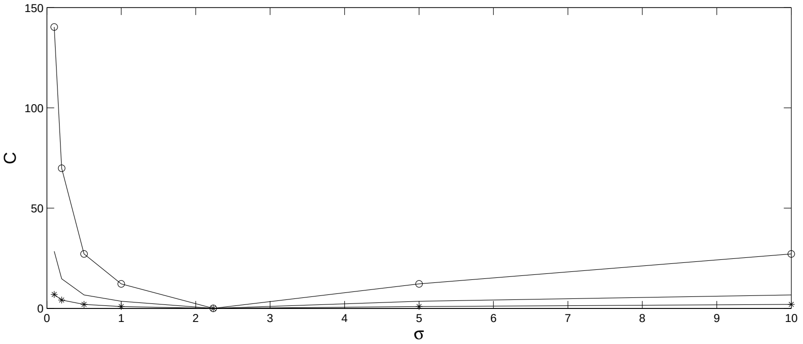

Figure I: Relative cost C versus σ. −○−, — and −∗− correspond to $\boldsymbol{\eta}$ multiplied by 0.2, 1 and 5 respectively.

when the signal level decreases and the discrepancy between σ and σ_0 increases. When $\sigma = \sigma_0$, $C = 0$.

F. Additional plots of standard deviations and absolute biases

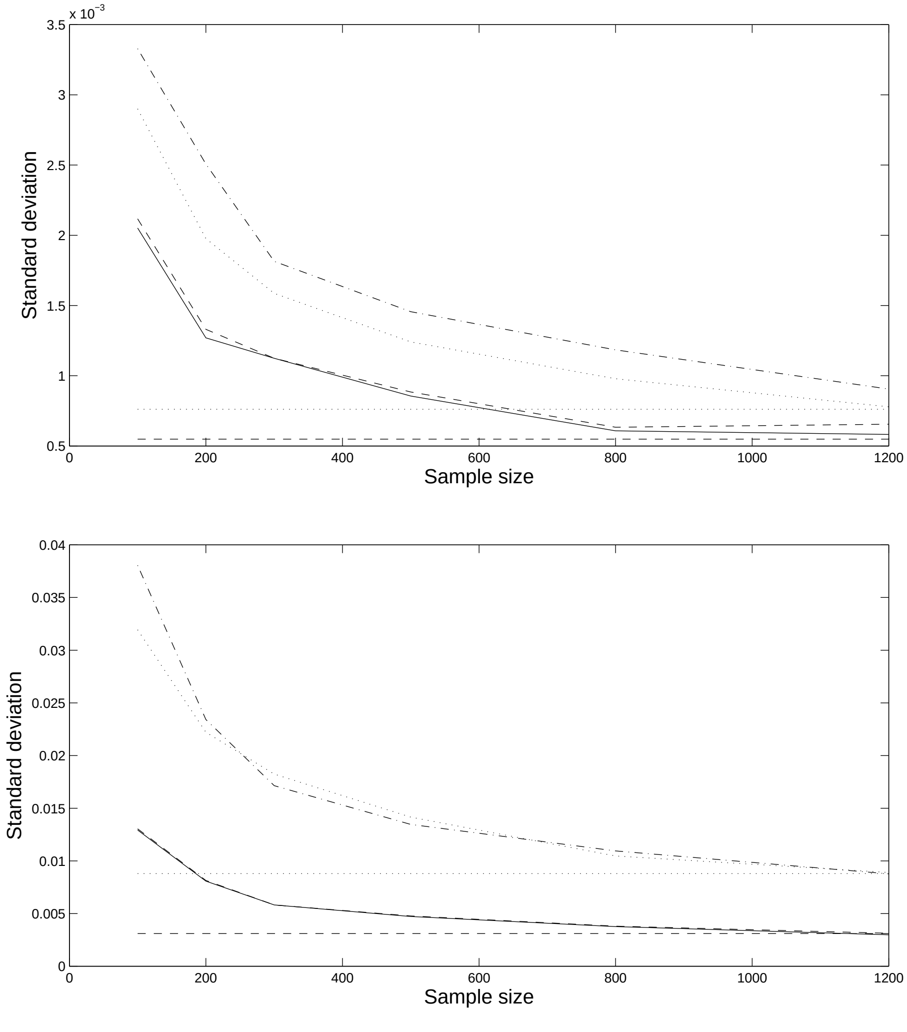

Figure II: Plots of standard deviations for the estimators in Figure 2: Comparison of SPE, SIMPLS, SPLS and OLS on estimation performance. The two horizontal lines mark the asymptotic standard deviations: Dashed: SPE; dotted: OLS. Other lines mark the sample standard deviations: Dash-dotted: SIMPLS; dotted: OLS; solid and dashed: SPE with starting values the true values and SPLS values. The solid and dashed lines overlap and are indistinguishable in the lower plot.

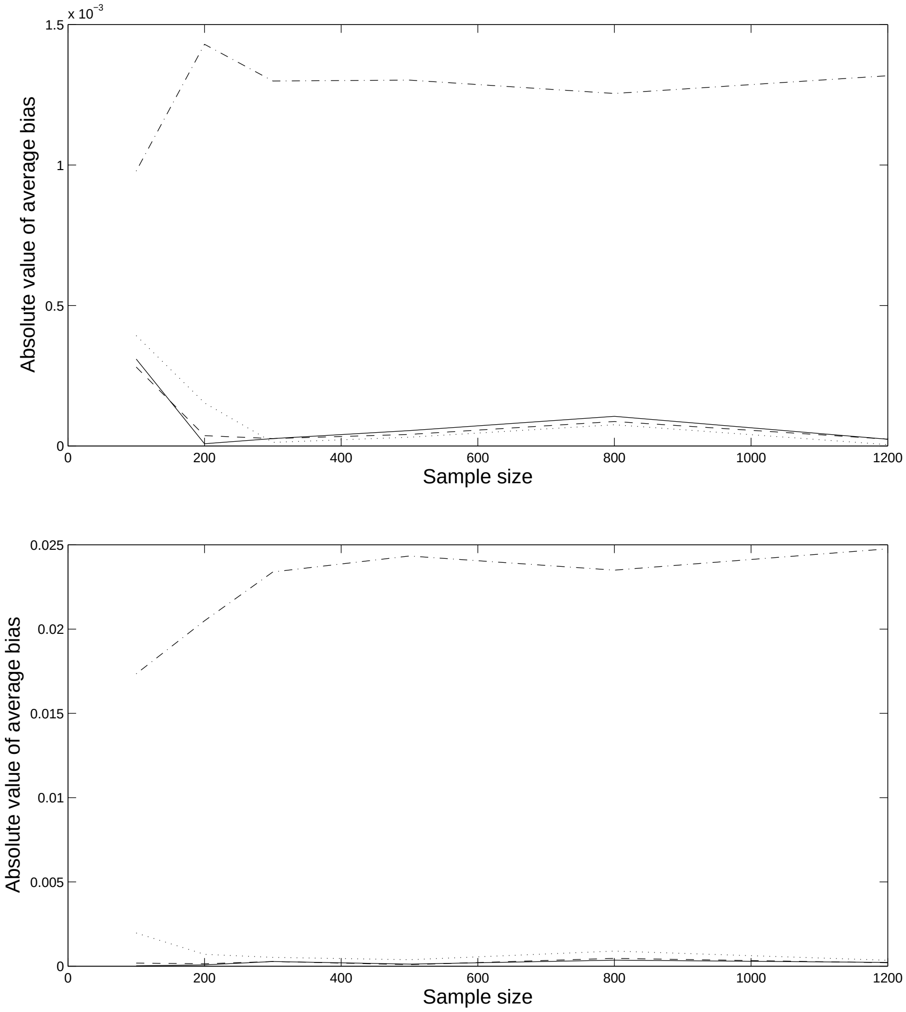

Figure III: Plots of absolute bias for the estimators in Figure 2: Comparison of the SPE, SIMPLS, SPLS and OLS estimator on estimation performance. Dash-dotted: SIMPLS. Dotted: OLS. The solid line and the dashed line overlap. They mark SPE with starting values using true value and SPLS.

References

Shapiro, A. (1986). Asymptotic Theory of Overparameterized Structural Models. *Journal of the American Statistical Association* **81**, 142–149.

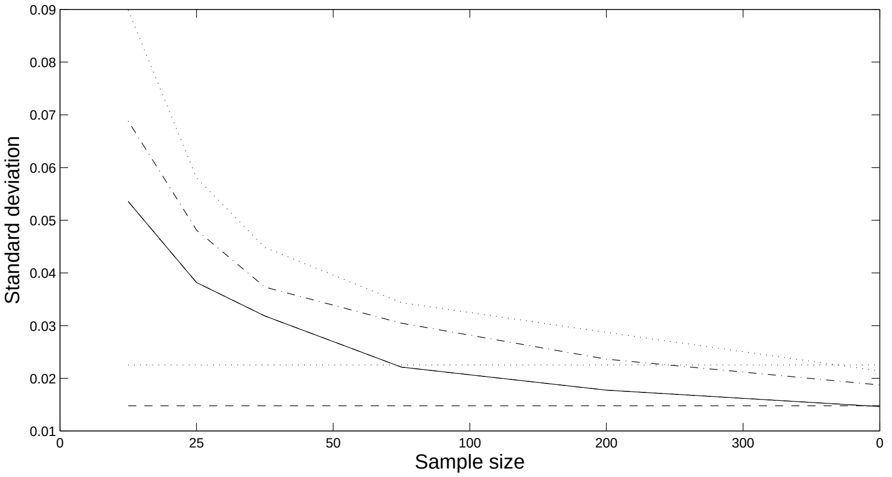

Figure IV: Plot of standard deviations for the estimators in Figure 4: Comparison of the SPE estimator, the scaled and ordinary SIMPLS estimators, and the OLS estimator when $\mathbf{\Lambda} = \mathbf{I}_p$. The line marks are the same as those in Figure II.

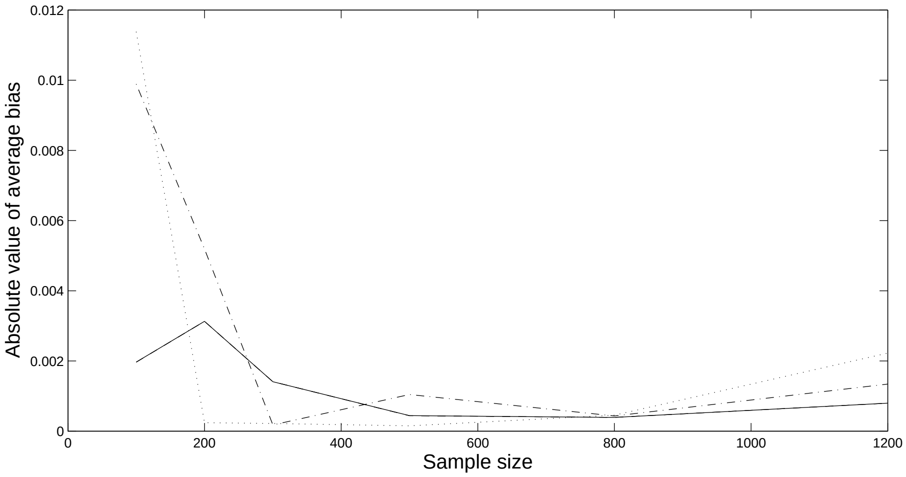

Figure V: Plot of absolute biases for estimators in Figure 4: Comparison of the SPE estimator, the scaled and ordinary SIMPLS estimators, and the OLS estimator when $\mathbf{\Lambda} = \mathbf{I}_p$. The line marks are the same as those in Figure III.

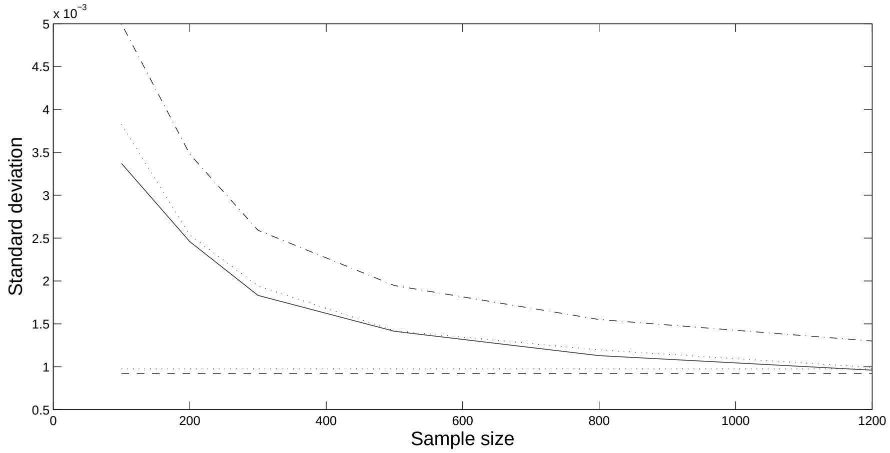

Figure VI: Plot of standard deviations for the estimators in Figure 5: Comparison of SPE, SIMPLS and OLS on estimation performance. The two horizontal lines mark the asymptotic standard deviations: Dashed: SPE; dotted: OLS. Other lines mark the sample standard deviations: Dash-dotted: SIMPLS; dotted: OLS; solid: SPE using the true values as starting values.

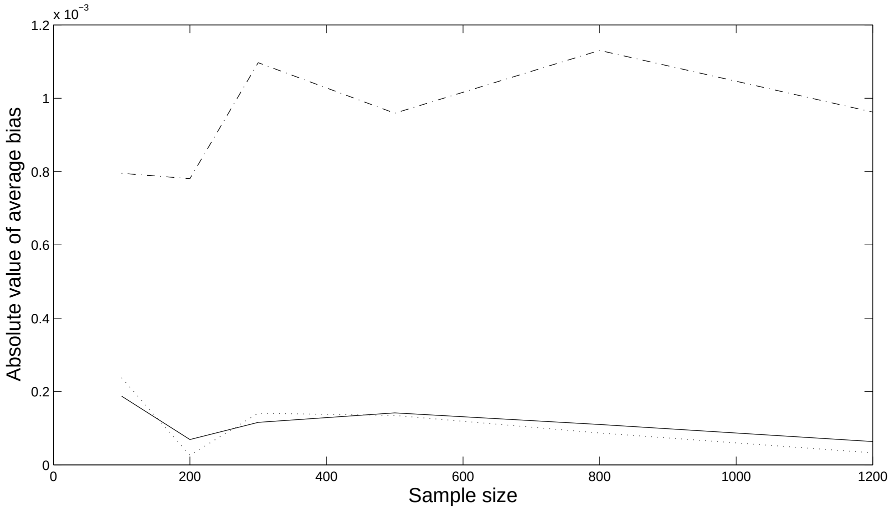

Figure VII: Plot of absolute biases for the estimators in Figure 5: Comparison of SPE, SIMPLS and OLS on estimation performance. Dash-dotted: SIMPLS; dotted: OLS; solid: SPE using the true values as starting values.