Stein-Rule Estimation and Generalized Shrinkage Methods for Forecasting Using Many Predictors

Eric Hillebrand
CREATES
Aarhus University, Denmark
ehillebrand@creates.au.dk

Tae-Hwy Lee
University of California, Riverside
Department of Economics
taelee@ucr.edu

August 2012

Abstract

We examine the Stein-rule shrinkage estimator for possible improvements in estimation and forecasting when there are many predictors in a linear time series model. We consider the Stein-rule estimator of Hill and Judge (1987) that shrinks the unrestricted unbiased OLS estimator towards a restricted biased principal component (PC) estimator. Since the Stein-rule estimator combines the OLS and PC estimators, it is a model-averaging estimator and produces a combined forecast. The conditions under which the improvement can be achieved depend on several unknown parameters that determine the degree of the Stein-rule shrinkage. We conduct Monte Carlo simulations to examine these parameter regions. The overall picture that emerges is that the Stein-rule shrinkage estimator can dominate both OLS and principal components estimators within an intermediate range of the signal-to-noise ratio. If the signal-to-noise ratio is low, the PC estimator is superior. If the signal-to-noise ratio is high, the OLS estimator is superior. In out-of-sample forecasting with AR(1) predictors, the Stein-rule shrinkage estimator can dominate both OLS and PC estimators when the predictors exhibit low persistence.

Keywords: Stein-rule, shrinkage, risk, variance-bias tradeoff, OLS, principal components.
JEL Classifications: C1, C2, C5

1 Introduction

Recent contributions to the forecasting literature consider many predictors in data-rich environments and principal components, such as Stock and Watson (2002, 2006, 2011), Bai (2003), Bai and Ng (2006, 2008), Bair et al. (2006), Hung and Lee (2010), Hillebrand et al. (2011), and Inoue and Kilian (2008), among others. In particular, Stock and Watson (2011) note that many forecasting models in this environment can be written in a unified framework called the shrinkage representation. Although the notion of the generalized shrinkage representation can be found in much earlier publications (e.g., Judge and Bock 1978), interest in shrinkage has been revived in the recent literature on out-of-sample forecasting.

The issue of forecasting using many predictors was discussed earlier in econometrics and statistics under the subject heading of ill-conditioned data or multicollinearity. In particular, Hill and Judge (1987) studied "improved prediction in the presence of multicollinearity." They examined possible improvements in estimation and forecasting when there are many predictors in a linear regression model. The Stein-rule estimator proposed in their paper shrinks the unrestricted unbiased OLS estimator towards a restricted biased principal component (PC) estimator.

Improvements are usually measured employing a risk function of the squared forecast error loss. While the asymptotic risk functions for the OLS and PC estimators are rather easily obtained, the risk of the Stein-rule is complicated as it depends on several unknown parameters and data-characteristics. It is not easy to understand conditions and situations under which improvements can be achieved. We conduct Monte Carlo simulations to shed light on the issue, both in-sample estimation and out-of-sample forecasting. In general, a key feature is that the desired improvement through Stein-rule shrinkage depends on the signal-to-noise ratio, which is affected by multiple determinants. The Stein-rule shrinkage estimator can dominate both OLS and PC estimators within an intermediate range of the signal-to-noise ratio. If the signal-to-noise ratio is low, the PC estimator tends to be superior. If the signal-to-noise ratio is high, the OLS estimator tends to be superior. In out-of-sample forecasting with AR(1) predictors, the Stein-rule shrinkage estimator can dominate both OLS and PC estimators when the predictors have low persistence.

Hill and Fomby (1992) examined the out-of-sample performance of a variety of biased estimation procedures such as ridge regression, principal component regression, and several Stein-like estimators. Their setup of evaluation was out-of-sample prediction in the sense that the out-of-sample data are different from the data used for parameter estimation, but not out-of-sample prediction in the context of the recent time series forecasting literature.

As the Stein-rule estimator of Hill and Judge (1987) combines OLS and PC estimators, it can be shown that it is a model-averaging estimator and thus produces a combined forecast. In fact, Hansen (2011) shows that the Stein-type shrinkage estimator is a Mallow-type combined estimator. Other papers have studied the relation between Stein-like shrinkage and forecast combinations. Fomby and Samanta (1991) use the Stein-rule for directly combining forecasts. Clark and McCracken (2009) examine the properties of combined forecasts of two nested models and note that their combined forecast is a Stein-type shrinkage forecast. Hence, the shrinkage principle provides insights not only into how to solve the issues of estimation in the presence of multicollinearity and forecasting using many predictors, but also how forecast combinations in the sense of Bates and Granger (1969) yield improvements.

The paper is organized as follows. Section 2 presents the shrinkage representation for forecasting using principal components. In Section 3 we consider the OLS and PC estimators and their asymptotic risk of the squared error loss. In Section 4, the Stein-rule shrinkage estimator that combines the OLS and PC estimators is presented. Section 5 and Section 6 present Monte Carlo analysis for in-sample and

out-of-sample performance of these three estimators – OLS, PC and the Stein-rule estimators. Finally, Section 7 provides some concluding remarks.

2 Shrinkage Representation

This section uses Stock and Watson's (2011) notation. Let the time series under study be denoted by y_t and let P_{it}, $i = 1, \ldots, K$, be a set of K orthonormal predictors such that $P'P/T = I_K$. These predictors can be thought of as the principal components of a possibly large data set X_{t-1}. The statistical model is

$$y_t = \delta' P_{t-1} + \varepsilon_t, \tag{1}$$

where $\delta \in \mathbb{R}^K$ is a parameter vector and ε_t is some error with mean zero and variance σ^2. Both y_t and P_t are assumed to have sample mean zero. Let $\tilde{y}_{T+1|T}$ be the forecast of y at time $T+1$ given information through time T. The theorems in Stock and Watson (2011) show that an array of forecasting methods, namely Normal Bayes, Bayesian Model Averaging, Empirical Bayes, and Bagging, have a *shrinkage representation*

$$\tilde{y}_{T+1|T} = \sum_{i=1}^{K} \psi(\kappa t_i) \hat{\delta}_i P_{iT} + o_P(1), \tag{2}$$

where $\hat{\delta}_i = T^{-1} \sum_{t=1}^{T} P_{i,t-1} y_t$ is the OLS estimator of δ_i, $t_i = \sqrt{T}\hat{\delta}_i/\hat{\sigma}$ is the t-statistic for $\hat{\delta}_i$, $\hat{\sigma}^2 = \sum_{t=1}^{T}(y_t - \hat{\delta}' P_{t-1})^2/(T-K)$ is the consistent estimator of σ^2, ψ is a function that is specific to a forecasting method, and κ is a constant that is specific to a forecasting method. For example, the shrinkage representation of the OLS estimator is $\psi(\kappa t_i) = 1$ for all i. A pre-test estimator has shrinkage representation $\psi(\kappa t_i) = \mathbb{1}_{\{|t_i| > t_c\}}$ for some critical value t_c. The principal components estimator that retains the first K_1 principal components and discards the others has shrinkage representation $\psi(\kappa t_i) = 1$ for $i \in \{1, \ldots, K_1\}$ and $\psi(\kappa t_i) = 0$ else. See also Judge and Bock (1978, p. 231) and Hill and Fomby (1992, p. 6) for a general representation of a family of minimax shrinkage estimators.

3 Principal Component Model

This section follows Hill and Judge (1987, 1990) and Hill and Fomby (1992), with adapted notation. Let the model in terms of the original predictor X be

$$y = X\beta + \varepsilon, \tag{3}$$

where y is a $T \times 1$ time series, X is a $T \times K$ matrix of K predictors, β is a $K \times 1$ parameter vector, and ε is a $T \times 1$ error time series with the conditional mean zero $\mathbb{E}(\varepsilon|X) = 0$ and conditional variance $\mathbb{E}(\varepsilon\varepsilon'|X) = \sigma^2 I_K$. Note that we do not assume normality of ε in this section while we generate it from the normal distribution in our simulation study in Sections 5 and 6. Our interest is to forecast y

when the number K of predictors in X is large. The location vector β is unknown and the objective is to estimate it by $\beta(y, X)$. We consider three estimators for β in this paper: (i) the ordinary least squares (OLS) estimator denoted $\hat{\beta}$, (ii) the principal component (PC) estimator denoted $\hat{\beta}^*$, and (iii) the Stein-like combined estimator of $\hat{\beta}$ and $\hat{\beta}^*$, which is to be denoted as $\tilde{\beta}$ in the next section. In this section we examine the sampling properties of $\hat{\beta}$ and $\hat{\beta}^*$ in terms of the asymptotic risk under the weighted squared error loss. In Sections 5 and 6 we compare them with the Stein-like combined estimator $\tilde{\beta}$.

The sampling performance of an estimator $\beta(y, X)$ is evaluated by its risk function, the expected weighted squared error loss with weight Q,

$$\text{Risk}\left(\beta, \beta(y, X), Q\right) = \mathbb{E}\left[(\beta(y, X) - \beta)' Q (\beta(y, X) - \beta)\right]. \tag{4}$$

As we will examine the performance of the Stein-like estimator in dynamic models for forecasting with weakly dependent time series, the predictor matrix X is treated as stochastic. Hence, the expectation in (4) is taken over the joint probability law of (y, X). In this section we compute the weighted quadratic risk with a weight $Q = X'X$, which gives the squared conditional prediction error risk. In Sections 5 and 6 we also consider a weight $Q = I_K$. The asymptotic risks of $\hat{\beta}$ and $\hat{\beta}^*$ are computed below based on the asymptotic covariances of $\hat{\beta}$ and $\hat{\beta}^*$.

The OLS estimator

$$\hat{\beta} = (X'X)^{-1} X'y, \tag{5}$$

conditional on X, has the asymptotic sampling property

$$\sqrt{T}\left(\hat{\beta} - \beta\right)\Big|_X \xrightarrow{d} \mathcal{N}\left(0,\, \sigma^2 (X'X)^{-1}\right). \tag{6}$$

The asymptotic quadratic risk weighted with $Q = X'X$ of the OLS estimator $\hat{\beta}$ is

$$\begin{aligned}
\text{Risk}\left(\beta, \hat{\beta}, X'X\right) &= \mathbb{E}\left\{\left(\hat{\beta} - \beta\right)' X'X \left(\hat{\beta} - \beta\right)\right\} \\
&= \text{tr}\ \mathbb{E}\left\{X'X \left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right\} \\
&= \text{tr}\ \mathbb{E}\left\{X'X \mathbb{E}\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)' | X\right]\right\} \\
&= \text{tr}\ \mathbb{E}\left\{X'X\ \sigma^2 (X'X)^{-1}\right\} \\
&= \text{tr}(\sigma^2 I_K) \\
&= K\sigma^2.
\end{aligned} \tag{7}$$

Since the bias $\mathbb{E}\left(\hat{\beta} - \beta | X\right) = 0$ conditional on X, the risk contains only a variance component.

Turning to the PC estimator $\hat{\beta}^*$, let V be the $K \times K$ matrix of eigenvectors of $X'X = T V \Lambda V'$, where Λ is the diagonal matrix of eigenvalues in descending order. Then, $V'V = I_K$ and

$$Y = X\beta + \varepsilon = XV\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}V'\beta + \varepsilon = P\delta + \varepsilon,\ P = XV\Lambda^{-\frac{1}{2}},\ \delta = \Lambda^{\frac{1}{2}}V'\beta. \tag{8}$$

This is the principal components regression model; P contains the principal components of X, and $\hat{\delta} = (P'P)^{-1} P'y = T^{-1}P'y$ can be estimated either from the principal components or as $\hat{\delta} = \Lambda^{\frac{1}{2}} V'\hat{\beta}$ from the OLS estimator of β. So far, the principal components model is equivalent to the original model. When X has a large degree of collinearity, the eigenvalues in Λ vary greatly in magnitude, and some are close to zero. Then, the number of components is decomposed into $K = K_1 + K_2$, where K_1 is the number of eigenvalues that are relatively large and K_2 is the number of eigenvalues that are relatively close to zero. The K_2 principal components that correspond to the small eigenvalues are discarded; the remaining K_1 principal components are kept. The model becomes

$$y = P\delta + \varepsilon = (P_1\ P_2) \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} + \varepsilon = P_1\delta_1 + P_2\delta_2 + \varepsilon, \tag{9}$$

$$= X(V_1\ V_2)\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}(V_1\ V_2)'\beta + \varepsilon = XV_1\Lambda_1^{-\frac{1}{2}}\Lambda_1^{\frac{1}{2}}V_1'\beta + XV_2\Lambda_2^{-\frac{1}{2}}\Lambda_2^{\frac{1}{2}}V_2'\beta + \varepsilon, \tag{10}$$

where Λ_1 and Λ_2 are the $K_1 \times K_1$ and $K_2 \times K_2$ diagonal matrices, respectively, that contain the corresponding eigenvalues, and $P_2\delta_2 = XV_2\Lambda_2^{-\frac{1}{2}}\Lambda_2^{\frac{1}{2}}V_2'\beta$ is deleted. Therefore, principal components regression with deleted components is equivalent to OLS estimation with the restriction

$$\delta_2 = R\beta = \Lambda_2^{\frac{1}{2}}V_2'\beta = 0, \tag{11}$$

where $R = \Lambda_2^{\frac{1}{2}}V_2'$ imposes K_2 linear restrictions on β. Note that $R = \Lambda_2^{\frac{1}{2}}V_2'$ is stochastic depending on X, and the risk of the restricted estimator is the expected loss with expectation taken over (y, X).

The principal components estimator of δ with K_2 deleted components, corresponding to the restrictions $\delta_2 = 0$, is

$$\hat{\delta}_1 = (P_1'P_1)^{-1}P_1'Y = T^{-1}P_1'Y. \tag{12}$$

The asymptotic distribution conditional on X is

$$\sqrt{T}\left(\hat{\delta}_1 - \delta_1\right)\Big|_X \xrightarrow{d} \mathcal{N}\left(0,\, \sigma^2 I_{K_1}\right). \tag{13}$$

The estimator $\hat{\delta}_1$ and setting $\delta_2 = 0$ result in the fit

$$y = P_1\hat{\delta}_1 + \hat{\varepsilon} = X\,V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1 + \hat{\varepsilon}, \tag{14}$$

and the principal components estimator of β is therefore

$$\hat{\beta}^* = V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1. \tag{15}$$

This is a special case of the restricted least squares (RLS) estimators explored by Mittelhammer (1985). Fomby, Hill, and Johnson (1978) present an optimality property of $\hat{\beta}^*$ that the trace of the asymptotic covariance matrix of $\hat{\beta}^*$ obtained by deleting K_2 principal components associated with the smallest eigenvalues is at least as small as that for any other RLS estimator with $J \leq K_2$ restrictions. This optimality is in terms of the asymptotic quadratic risk weighted with $Q = I_K$, i.e., Risk $\left(\beta, \hat{\beta}^*, I_K\right)$.

For forecasting, it is interesting to examine the asymptotic quadratic risk weighted with $Q = X'X$ of the PC estimator $\hat{\beta}^*$.

$$\begin{aligned}
\text{Risk}\left(\beta, \hat{\beta}^*, X'X\right) &= \mathbb{E}\left(\hat{\beta}^* - \beta\right)' X'X \left(\hat{\beta}^* - \beta\right) \\
&= \mathbb{E}\left(V_1 \Lambda_1^{-\frac{1}{2}} \hat{\delta}_1 - \beta\right)' (T\, V \Lambda V') \left(V_1 \Lambda_1^{-\frac{1}{2}} \hat{\delta}_1 - \beta\right) \\
&= T\mathbb{E}\left(V_1 \Lambda_1^{-\frac{1}{2}} \hat{\delta}_1 - \beta\right)' \left(V \Lambda^{1/2} \Lambda^{1/2} V'\right) \left(V_1 \Lambda_1^{-\frac{1}{2}} \hat{\delta}_1 - \beta\right) \\
&= T\mathbb{E}\left(\hat{\delta}_1' \Lambda_1^{-\frac{1}{2}} V_1' V \Lambda^{1/2} - \beta' V \Lambda^{1/2}\right) \left(\Lambda^{1/2} V' V_1 \Lambda_1^{-\frac{1}{2}} \hat{\delta}_1 - \Lambda^{1/2} V' \beta\right) \\
&= T\mathbb{E}\left(\hat{\delta}_1' \left[\begin{array}{cc} I_{K_1} & \mathbf{0} \end{array}\right] - \delta'\right) \left(\left[\begin{array}{c} I_{K_1} \\ \mathbf{0} \end{array}\right] \hat{\delta}_1 - \delta\right) \\
&= T\mathbb{E}\left(\hat{\delta}_1 - \delta_1\right)' \left(\hat{\delta}_1 - \delta_1\right) + T\delta_2'\delta_2 \\
&= T\,\text{tr}\ \mathbb{E}\left(\hat{\delta}_1 - \delta_1\right) \left(\hat{\delta}_1 - \delta_1\right)' + T\delta_2'\delta_2 \\
&= T\,\text{tr}\ \left(T^{-1}\sigma^2 I_{K_1}\right) + T\delta_2'\delta_2 \\
&= K_1\sigma^2 + T\delta_2'\delta_2
\end{aligned} \tag{16}$$

where the first term corresponds to the variance term which declines as K_1 decreases and the second term corresponds to the bias term. The second to the last equality follows from (13).

To compare the asymptotic risks of the OLS $\hat{\beta}$ estimator and the PC estimator $\hat{\beta}^*$, look at the risk difference

$$\text{Risk}\left(\beta, \hat{\beta}, X'X\right) - \text{Risk}\left(\beta, \hat{\beta}^*, X'X\right) = K\sigma^2 - \left(K_1\sigma^2 + T\delta_2'\delta_2\right) = K_2\sigma^2 - T\delta_2'\delta_2,$$

which is positive when $\delta_2'\delta_2$ is small. This is the case if the restriction in (11) is reasonable. In that case the OLS estimator $\hat{\beta}$ is dominated by the PC estimator $\hat{\beta}^*$.

4 Stein-Rule Estimator

Hill and Judge (1987, 1990) propose a Stein-rule estimator $\tilde{\beta}$ that shrinks the standard OLS estimator $\hat{\beta}$ towards the principal components estimator $\hat{\beta}^*$:

$$\begin{aligned}
\tilde{\beta} &= \hat{\beta}^* + \left(1 - \frac{a\hat{\sigma}^2(T-K)}{\hat{\beta}' R' (R(X'X)^{-1}R')^{-1} R\hat{\beta}}\right) (\hat{\beta} - \hat{\beta}^*) \\
&= \hat{\beta}^* + \psi(\hat{\beta} - \hat{\beta}^*) \\
&= \psi\hat{\beta} + (1-\psi)\hat{\beta}^*
\end{aligned} \tag{17}$$

where a is a constant, R is defined from (11), and the Stein coefficient ψ is the shrinkage from the OLS estimator $\hat{\beta}$ to the PC estimator $\hat{\beta}^*$. Using $R = \Lambda_2^{\frac{1}{2}} V_2'$ and $\hat{\beta} = V\Lambda^{-\frac{1}{2}}\hat{\delta}$, we obtain

$$\tilde{\beta} = V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1 + \left(1 - \frac{1}{T}\frac{a\hat{\sigma}^2(T-K)}{\hat{\delta}'\Lambda^{-\frac{1}{2}}V'V_2\Lambda_2^{\frac{1}{2}}(\Lambda_2^{\frac{1}{2}}V_2'V\Lambda^{-1}V'V_2\Lambda_2^{\frac{1}{2}})^{-1}\Lambda_2^{\frac{1}{2}}V_2'V\Lambda^{-\frac{1}{2}}\hat{\delta}}\right)(V\Lambda^{-\frac{1}{2}}\hat{\delta} - V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1). \tag{18}$$

Using that

$$V'V_2 = \begin{bmatrix} \mathbf{0} \\ I_{K_2} \end{bmatrix}, \quad V_2'V = [\mathbf{0}\ I_{K_2}], \text{ and } (X'X)^{-1} = T^{-1}V\Lambda^{-1}V',$$

where I_{K_2} is the $K_2 \times K_2$ identity matrix, we obtain that

$$\begin{aligned}\tilde{\beta} &= \left(1 - \frac{1}{T}\frac{a\hat{\sigma}^2(T-K)}{\hat{\delta}_2'\hat{\delta}_2}\right)(V\Lambda^{-\frac{1}{2}}\hat{\delta} - V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1) + V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1, \\ &= V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1 + \psi(V\Lambda^{-\frac{1}{2}}\hat{\delta} - V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1). \end{aligned} \tag{19}$$

Further rearrangement yields the expression

$$\tilde{\beta} = V_1\Lambda_1^{-\frac{1}{2}}\hat{\delta}_1 + \psi V_2\Lambda_2^{-\frac{1}{2}}\hat{\delta}_2, \tag{20}$$

for the Stein-rule estimator, from which its shrinkage representation can now be read. Since the individual t-statistics of the principal components are given by $t_i = \sqrt{T}\hat{\delta}_{2,i}/\hat{\sigma}$, the coefficient of the K_2 terms in $\hat{\delta}$ corresponding to the discarded principal components can be written

$$1 - \frac{1}{T}\frac{a\hat{\sigma}^2(T-K)}{\hat{\delta}_2'\hat{\delta}_2} = 1 - \frac{a(T-K)}{\sum_{K_1+1}^{K} t_i^2} = 1 - \frac{a(T-K)}{K_2 F_{K_2,T-K}}, \tag{21}$$

where $F_{K_2,T-K} = \sum_{K_1+1}^{K} t_i^2/K_2$ is the test statistic for $H_0 : \delta_2 = 0$, the restriction of Equation (11). Note that

$$F_{K_2,T-K} = \frac{\sum_{K_1+1}^{K} t_i^2}{K_2} = \frac{T\hat{\delta}_2'\hat{\delta}_2/K_2}{\hat{\sigma}^2} = \frac{\text{signal from } K_2 \text{ discarded variables}}{\text{noise}}.$$

The Stein coefficient function ψ in the shrinkage representation of Section 2 is given by

$$\psi_i = \begin{cases} 1, & i \in \{1, \ldots, K_1\}, \\ \psi, & i \in \{K_1+1, \ldots, K\}. \end{cases} \tag{22}$$

The asymptotic quadratic risk weighted with $Q = X'X$ for the Stein estimator $\tilde{\beta}$

$$\text{Risk}\left(\beta, \tilde{\beta}, X'X\right) = \mathbb{E}\left[\left(\tilde{\beta} - \beta\right)' X'X \left(\tilde{\beta} - \beta\right)\right] \tag{23}$$

can be calculated here but it is rather complicated as it depends on parameters such as β, σ, a, and on data characteristics such as T, K, K_1, X (with X determining Λ, V). Hence, we use Monte Carlo analysis in the next two sections to examine the risk of $\tilde{\beta}$ in comparison with those of $\hat{\beta}$ and $\hat{\beta}^*$.

5 In-Sample Performance of Stein-Rule Shrinkage Estimator

We conduct Monte Carlo analysis to compare the risk of $\tilde{\beta}$ with those of $\hat{\beta}$ and $\hat{\beta}^*$. The risk of the Stein-rule estimator depends on β, σ, a, T, K, K_1, X. For the risk comparisons we fix $T = 200$ and $K = 50$ while we vary β, σ, a, K_1, and X.

5.1 Simulation Design

The elementary model to be studied is the linear equation

$$y = X\beta + \varepsilon, \tag{24}$$

where y is a $T \times 1$ vector, X is a $T \times K$ matrix of regressors, ε is a $T \times 1$ random vector drawn from $\mathcal{N}(0, \sigma^2)$ distribution, $\beta \in \mathbb{R}^{K\times 1}$, and $\sigma \in \mathbb{R}_+$.

We compare the performance of the Stein-rule estimator in-sample with the standard OLS estimator and the principal components estimator and employ the following simulation design. We draw a matrix X_0 of $\mathcal{N}(0, 1)$ random variables of dimensions $T \times K$, $T = 200$, $K = 50$. We aim to impose different eigenvalue structures on the regressor matrix X in the spirit of Hill and Judge (1987). To this end, we singular-value decompose X_0 into

$$X_0 = U\Lambda_0^{\frac{1}{2}}V'$$

and discard the diagonal matrix $\Lambda_0^{\frac{1}{2}}$. The regressor matrix X is then constructed as

$$X = U\Lambda^{\frac{1}{2}}V', \tag{25}$$

where Λ is constructed according to three different scenarios.

- The singular values are constant.
$$\Lambda^{\frac{1}{2}} = \text{diag}(2, \ldots, 2). \tag{26}$$
- The singular values are linearly decreasing from 5 to 1.
- The singular values are exponentially decreasing from 5 to 1.
$$\text{diag}(\Lambda^{\frac{1}{2}}) = 1 + 4e^{-0.10k}, \quad k = 1, \ldots, K. \tag{27}$$

In the data-generating process, we consider different scenarios for the variance σ^2 of the error process. In particular, we set

$$\sigma \in \{1, 3, 5, 7\}. \tag{28}$$

The data-generating parameter vector β is set to

$$\beta = \left[\frac{L}{K}\right]_{k\in\{1,\ldots,K\}}, \tag{29}$$

such that its direction in parameter space is $(1/K)_k$ and its length is L. We consider different scenarios for the length L of the vector, in particular,

$$L \in \{0,\, 1,\, 2,\, 3\}. \tag{30}$$

Table 1 lists the population-R^2 for the different resulting scenarios, where

$$\text{population-}R^2 = \frac{\beta'\mathbb{E}(X'X)\beta}{\beta'\mathbb{E}(X'X)\beta + T\sigma^2}.$$

Our simulation design considers only a limited region of the space of simulation design parameters (T, K, L, σ, Λ). Estimating a response surface for a larger region could give some more indication on the range of data-sets where gains from Stein-rule estimation can be expected. This is left for future research.

Table 1: Population R^2 for the different simulation scenarios.

Eigenvalues	constant				linear				exponential			
L / σ^2	1	3	5	7	1	3	5	7	1	3	5	7
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0.020	0.007	0.004	0.003	0.049	0.017	0.010	0.007	0.019	0.007	0.004	0.003
2	0.074	0.026	0.016	0.011	0.172	0.065	0.040	0.029	0.073	0.026	0.016	0.011
3	0.153	0.057	0.035	0.025	0.319	0.135	0.086	0.063	0.151	0.056	0.034	0.025

The performance of the estimators is measured in terms of their risk. The general risk function considered is

$$\text{Risk}\left(\beta, \beta(y,X), Q\right) = \mathbb{E}\left[\left(\beta\left(y,X\right)-\beta\right)' Q \left(\beta\left(y,X\right)-\beta\right)\right],$$

as shown in (4). We study the particular case where $Q = I$, which results in the standard mean squared error considered in James and Stein (1961) and Judge and Bock (1978) and the second case where $Q = X'X$ as considered in Judge and Bock (1978), Hill and Judge (1987, 1990), and Hill and Fomby (1992). This risk measure can be interpreted as the square of the distance $(X\beta - X\hat{\beta})$ of the fitted value from the signal part of y.

There are a few estimator-specific settings to consider as well, in particular the number K_1 of principal components for the principal components estimator and the value of the parameter a in the Stein-rule shrinkage estimator. We consider $K_1 \in \{1,\, 5,\, 10,\, 20\}$ and $a \in (0,\, 1)$. The two numbers

interact through the bounds for a given in Judge and Bock (1978, p. 193) and Hill and Judge (1987, p. 87):

$$0 \leq a \leq \frac{2(K - K_1 - 2)}{T - K + 2}. \tag{31}$$

Here, for $K_1 \in \{1, 5, 10, 20\}$, we obtain

$$0 \leq a \leq 0.62, 0.57, 0.50, 0.37,$$

so that we expect the region for a in which the Stein-rule shrinkage estimator performs better than OLS and PC estimators to move towards the origin as the number of components increases.

5.2 Choosing the Number of Principal Components

Selecting the number of principal components is a problem that has spawned a large literature (see, for example, Anderson 2003, Bai & Ng 2002, Hallin & Liska 2007, and Onatski 2009). In this paper, we restrict ourselves to studying the behavior of the Stein-rule estimator for a set of number of components, including the one-factor model, few components ($K_1 = 5$), a moderate number ($K_1 = 10$), and many factors ($K_1 = 20$). Recall that the number of regressors is $K = 50$.

Figures 1 to 3 show the risk of the three estimators, Stein-rule shrinkage, OLS, and PC, as functions of the parameter a of the Stein-rule shrinkage estimator. Since the OLS and PC estimators do not depend on this parameter, they are constants in the graphs. The risk of the OLS estimator is depicted by a dotted line; the risk of the PC estimator is shown as a dashed line. The risk of the Stein-rule shrinkage estimator is shown as connected dots. The left panel of four plots in each figure shows the MSE risk ($Q = I$); the right panel of four plots shows the risk for $Q = X'X$. The four plots show the different scenarios for the number $K_1 \in \{1, 5, 10, 20\}$ of principal components. Each figure shows a different singular value scenario, Figure 1 shows the case of constant singular values equal to two; Figure 2 shows the case of linearly decreasing singular values, and Figure 3 shows the case of exponentially decreasing singular values.

The graphs show that the risk of the Stein-rule follows a parabola in a, which indicates that there is an optimal a, at least in the simulation scenarios considered. Unlike in the case of the original James and Stein (1961) estimator, this optimal a is not analytically known at this point. The minimum of the parabola is moving inwards toward the origin as the number K_1 of components increases, as expected. For the scenarios where the singular values are constant and where they are linearly decreasing, the OLS estimator performs generally better than the PC estimator. For exponentially decreasing singular values, the PC estimator often performs better than the OLS estimator. The Stein-rule shrinkage estimator has a greater relative advantage over PC and OLS estimators for small number of components ($K_1 = 1, 5$). For larger K_1, the performance of the Stein-rule shrinkage estimator approaches that of the relatively better estimator among OLS and PC. Note that the singular value scenarios considered in this paper

do not include values close to zero as in Hill and Judge (1987). We found that for most scenarios of this nature, where a strong degree of multicollinearity is present, the principal components estimator performs better than the Stein-rule shrinkage estimator.

5.3 Different Variance Scenarios

Figures 4 through 6 report the performance of the estimators for different noise levels

$$\sigma \in \{1, 3, 5, 7\}.$$

The organization of the graphs is the same as described in Section 5.2. Again, the risk of the Stein-rule shrinkage estimator describes a parabola in a, indicating the existence of an optimal parameter value. For low values of variance, OLS performs better than principal components, and as the noise level increases, the PC estimator outperforms OLS. The Stein-rule shrinkage estimator can outperform both OLS and PC estimators within an intermediate noise range.

5.4 Different Lengths of the Parameter Vector β

Figures 7 through 9 display the performance of the estimators for different lengths L of the parameter vector $\beta = L/K$. The four plots of each panel show the risks of the estimators for

$$L \in \{0, 1, 2, 3\}.$$

The organization of the graphs is the same as described in Section 5.2. If $L = 0$, that is, there is no signal in y, the PC estimator outperforms both OLS and the Stein-rule shrinkage estimators. For large values of L, OLS performs better than the other estimators. On an intermediate range, the Stein-rule shrinkage estimator can outperform both other estimators.

Recall from Equation (11) that $\delta_2 = \Lambda_2^{\frac{1}{2}} V_2' \beta$ where $\beta = \left[\frac{L}{K}\right]$. Hence, the length L for β determines the length of δ_2. Because $T\delta_2'\delta_2$ is the second term in the asymptotic risk of the PC estimator corresponding to the bias due to the omission of the K_2 principal components, as shown in (16), a large value of L increases the risk of the PC estimator compared to the risk of the OLS estimator.

For the Stein-rule estimator, a large value of L increases $T\delta_2'\delta_2$, which in turn will increase the F statistic defined from (21)

$$F_{K_2,T-K} = \frac{T\hat{\delta}_2'\hat{\delta}_2/K_2}{\hat{\sigma}^2}$$

and hence increases the Stein-rule coefficient ψ and thus reduces the shrinkage from the OLS estimator $\hat{\beta}$ to the PC estimator $\hat{\beta}^*$.

6 Out-of-Sample Performance of Stein-Rule Shrinkage Estimator

6.1 Simulation Design

We assess the out-of-sample (OOS) performance of the Stein-rule shrinkage estimator in two different simulation setups. One is exactly the same as described in Section 5.1, only that the forecast performance on $T_2 = 100$ out-of-sample observations is evaluated. The two risk functions considered for the OOS comparison are the mean squared forecast error (MSFE)

$$\text{MSFE}\left(\beta(y,\, X)\right) = \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)], \tag{32}$$

where $\hat{y} = X\beta(y, X)$, and the squared signal-to-prediction distance as considered in (4) with $Q = X'X$,

$$\text{Risk}(\beta,\, \beta(y, X),\, X'X) = \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]. \tag{33}$$

The second simulation environment that we study has AR(1) time series in the columns of the regressor matrix X. That is,

$$X = [\{x_{1,t}\}\, \{x_{2,t}\}\, \ldots \{x_{K,t}\}]_{t\in\{1,\ldots,T\}}, \tag{34}$$

and the individual columns follow

$$x_{k,t} = \phi x_{k,t-1} + \sigma_{X,k}\xi_{t,k}, \quad k = 1, \ldots, K,\, \xi_{t,k} \sim \mathcal{N}(0, 1). \tag{35}$$

The standard deviations of the AR(1) processes in the columns of X are chosen to correspond to the exponentially decaying sequence employed in Equation (27):

$$\sqrt{\text{Var } x_{k,t}} = \frac{\sigma_{X,k}}{\sqrt{1-\phi^2}} = 1 + 4e^{-0.10k},\ k = 1,\, \ldots,\, K. \tag{36}$$

Thus, $\sigma_{X,k} = \sigma_{X,k}(\phi) = (1{+}4e^{-0.10k})\sqrt{1-\phi^2}$. Varying ϕ replaces the variance dimension considered in the in-sample study. The standard deviation of the noise in y is set to $\sigma = 3$ and $T_2 = 100$ out-of-sample observations are evaluated.

Note that principal components are linear combinations of the columns of X,

$$P_1 = XV_1\Lambda_1^{-\frac{1}{2}}, \tag{37}$$

and therefore the individual components are, with some coefficients $w_{k,j}$ determined by V_1 and Λ_1,

$$\begin{aligned} P_{j,t} &= \sum_{k=1}^{K} w_{k,j}x_{k,t} = \phi\sum_{k=1}^{K} w_{k,j}x_{k,t-1} + \sum_{k=1}^{K} w_{k,j}\sigma_{X,k}\xi_{t,k}, \\ &= \phi P_{j,t-1} + \eta_{j,t}, \end{aligned}$$

where $\eta_{j,t} = \sum_{k=1}^{K} w_{k,j}\sigma_{X,k}\xi_{t,k}$. As long as the AR(1) parameter ϕ is the same across all columns of X, the principal components will themselves be AR(1) processes with the same decorrelation length as the individual columns. If different ϕ_k are chosen across the columns, the principal components will be linear combinations of AR(1) processes with different persistence parameters, which can lead to long memory behavior of the components, as described in Granger (1980).

6.2 Choosing the Number of Principal Components

Figures 10 and 11 display the out-of-sample performance of the estimators for different numbers of principal components

$$K_1 \in \{1,\ 5,\ 10,\ 20\}.$$

The organization of the graphs is similar to the one described in Section 5.2. Instead of different singular value scenarios, two different simulation designs are considered. Figure 10 shows the case where the regressor matrix X is drawn from independent $\mathcal{N}(0,1)$ distributions; Figure 11 shows the case where the regressors are AR(1) time series. Unlike in the in-sample study, here the relative performance of PC and OLS estimators changes with the number K_1 of principal components. For small numbers, OLS performs better than PC, and for large K_1, PC performs better than OLS. The Stein-rule shrinkage estimator dominates for up to ten components. There is no obvious difference between the i.i.d. and the AR(1) simulation scenarios.

6.3 Different Variance Scenarios

Figure 12 shows the performance of the estimator when X is drawn from an $\mathcal{N}(0,1)$ distribution. Similar to the in-sample study, OLS performs best for low noise levels and PC performs best for high noise levels. The Stein-rule shrinkage estimator can outperform both in an intermediate noise range.

Figure 13 shows the performance for the estimators when the columns of X follow AR(1) dynamics. The four plots in each panel show the situation for different values $\phi \in \{0.30,\ 0.50,\ 0.90,\ 0.99\}$ of the AR-parameter. The standard deviation of the error in the AR model is then set through $\sigma_{X,k}(\phi) = (1+4e^{-0.10k})\sqrt{1-\phi^2}$ such that the standard deviation of the column follows Equation (36). The figure shows that the Stein-rule shrinkage estimator outperforms OLS and PC estimators in low persistence scenarios ($\phi = 0.30,\ 0.50$), whereas in high persistence scenarios ($\phi = 0.90,\ 0.99$) the PC estimator outperforms both Stein-rule and OLS. The relative performance of OLS and PC estimators also changes with persistence: In low persistence scenarios, OLS performs better than PC, and vice versa for high persistence.

6.4 Different Lengths of the Parameter Vector β

Figures 14 and 15 show the performance of the estimators for different lengths $L \in \{0, 1, 2, 3\}$ of the parameter vector. As in the in-sample case, when $L = 0$, PC performs best. For $L = 1$, OLS performs better than PC, but both are dominated by the Stein-rule shrinkage estimator. For higher values of L, OLS performs best among all three estimators. This holds true for both simulation environments, i.i.d. regressors and AR(1) regressors.

7 Concluding Remarks

In this paper, we have shown that the Stein-rule shrinkage estimator that shrinks the OLS estimator towards the PC estimator, as proposed in Hill and Judge (1987, 1990), can be represented as a shrinkage estimator for a forecasting model as proposed in Stock and Watson (2011). We examined the performance of the estimator in a variety of simulation environments, both in-sample and out-of-sample. The overall picture that emerges is that the Stein-rule shrinkage estimator can dominate both OLS and principal components estimators within an intermediate range of the signal-to-noise ratio. If the noise level is high (high variance of noise terms) or if the signal is low (short parameter vector), the principal components estimator is superior. If the noise level is low (low variance of noise terms) or if the signal is high (long parameter vector), OLS is superior. In out-of-sample simulations with AR(1) regressors, the Stein-rule shrinkage estimator can dominate both OLS and principal components estimators in low persistence situations.

Acknowledgments

We would like to thank the participants of the Advances in Econometrics conference in Baton Rouge in March 2012 for their helpful comments, in particular Carter Hill, Tom Fomby, Stan Johnson, Mike McCracken, and Lee Adkins. We would also like to thank the organizers of the conference, in particular Dek Terrell, for a job superbly done and for many useful comments on this paper. The usual disclaimer applies.

References

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis.* 3rd edn. Hoboken, NJ: Wiley.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1), 135–171.

Bai, J., & Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1), 191–221.

Bai, J., & Ng, S. (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4), 1133–1150.

Bai, J., & Ng, S. (2008). Forecasting Economic Time Series Using Targeted Predictors. *Journal of Econometrics*, 146, 304–317.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101(473), 119–137.

Bates, J.M., & Granger, C.W.J. (1969). The Combination of Forecasts. *Operations Research Quarterly*, 20, 451–468.

Clark, T.E., & McCracken, M.W. (2009). Combining Forecasts from Nested Models. *Oxford Bulletin of Economics and Statistics*, 71(3), 303–329.

Fomby, T.B., Hill, R.C., & Johnson, S.R. (1978). An Optimal Property of Principal Components in the Context of Restricted Least Squares. *Journal of the American Statistical Association*, 73(361), 191–193.

Fomby, T.B., & Samanta, S.K. (1991). Application of Stein Rules to Combination Forecasting. *Journal of Business and Economic Statistics*, 9(4), 391–407.

Granger, C.W.J. (1980). Long Memory Relationships and the Aggregation of Dynamic Models. *Journal of Econometrics*, 14, 227–238.

Hansen, B. (2011). Efficient Shrinkage in Parametric Models. Manuscript, University of Wisconsin, Madison.

Hill, R.C., & Fomby, T.B. (1992). The Effect of Extrapolation on Minimax Stein-Rule Prediction. In W. Griffiths, H. Lütkepohl, & M.E. Bock (Eds.), *Readings in Econometric Theory and Practice, A Volume in Honor of George G. Judge* (pp. 3–32). Amsterdam: North-Holland.

Hill, R.C., & Judge, G.G. (1987). Improved Prediction in the Presence of Multicollinearity. *Journal of Econometrics*, 35, 83–100.

Hill, R.C., & Judge, G.G. (1990). Improved Estimation under Collinearity and Squared Error Loss. *Journal of Multivariate Analysis*, 32, 296–312.

Hillebrand, E., Huang, H., Lee, T.-H., & Li, C. (2011). Using the Yield Curve in Forecasting Output Growth and Inflation. Manuscript, Aarhus University and UC Riverside.

Huang, H., & Lee, T.-H. (2010). To Combine Forecasts or To Combine Information? *Econometric Reviews*, 29, 534–570.

Inoue, A., & Kilian, L. (2008). How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation. *Journal of the American Statistical Association*, 103(482), 511–522.

James, W., & Stein, C.M. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.

Judge, G.G., & Bock, M.E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Hallin, M. & Liska, R. (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association*, 102(478), 603–617.

Mittelhammer, R.C. (1985). Quadratic Risk Domination of Restricted Least Squares Estimators via Stein-Rules Auxiliary Constraints. *Journal of Econometrics*, 29, 289–304.

Onatski, A. (2009). Testing Hypotheses About the Number of Factors in Large Factor Models. *Econometrica*, 77(5), 1447–1479.

Stock, J.H., & Watson, M.W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97, 1167–1179.

Stock, J.H., & Watson, M.W. (2006). Forecasting with Many Predictors. In: G. Elliott, C.W.J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume 1* (pp. 515–554). Amsterdam: North Holland.

Stock, J.H., & Watson, M.W. (2011). Generalized Shrinkage Methods for Forecasting Using Many Predictors. Manuscript, Harvard University and Princeton University.

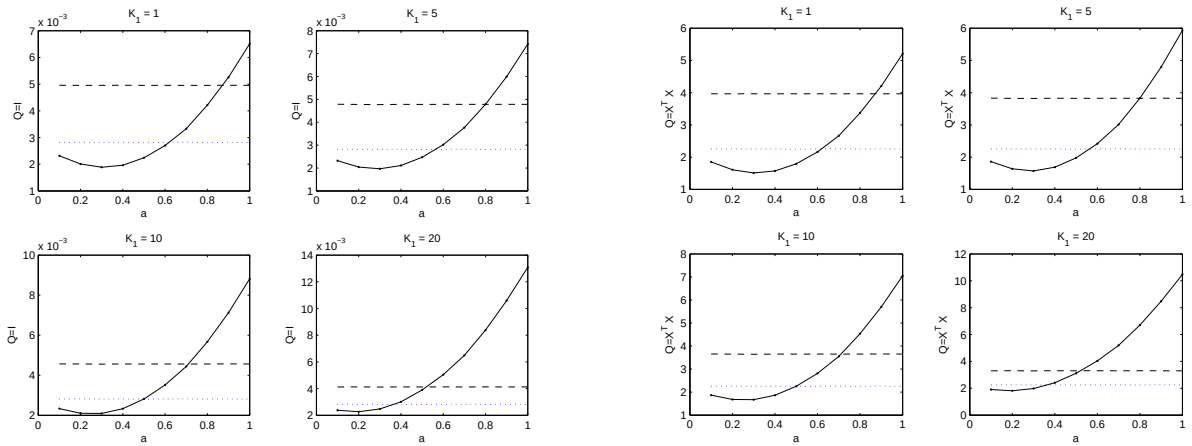

Figure 1: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are constant and equal to two. Other parameters are set to $L = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

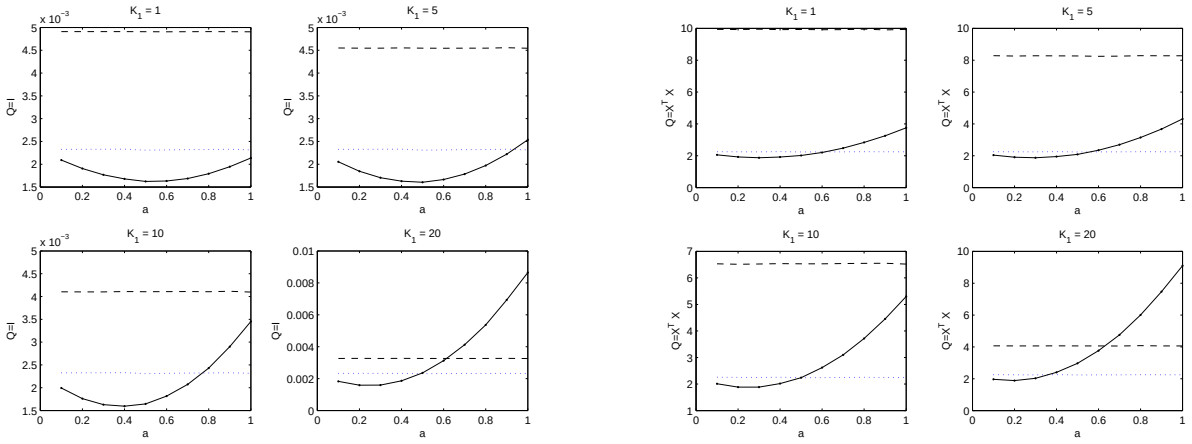

Figure 2: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are linearly decreasing from 5 to 1. Other parameters are set to $L = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

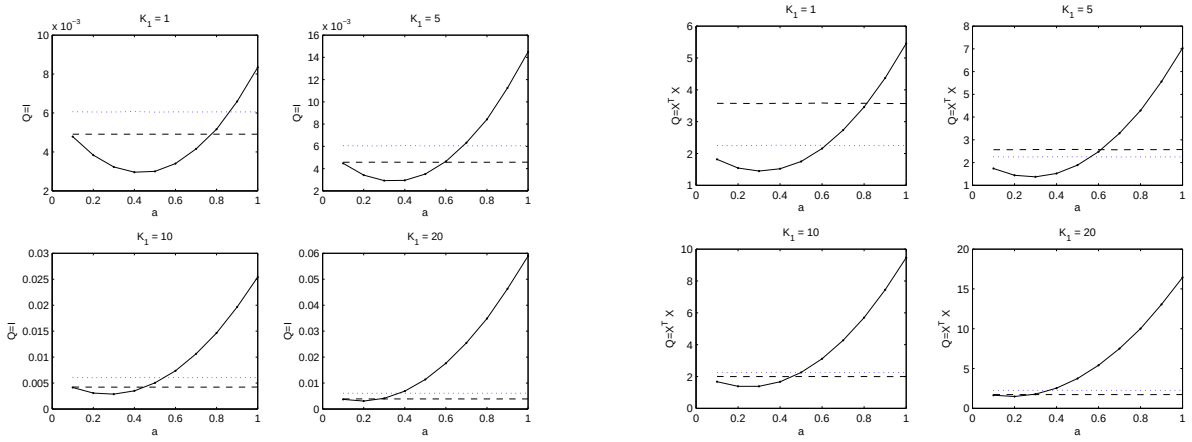

Figure 3: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $L = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

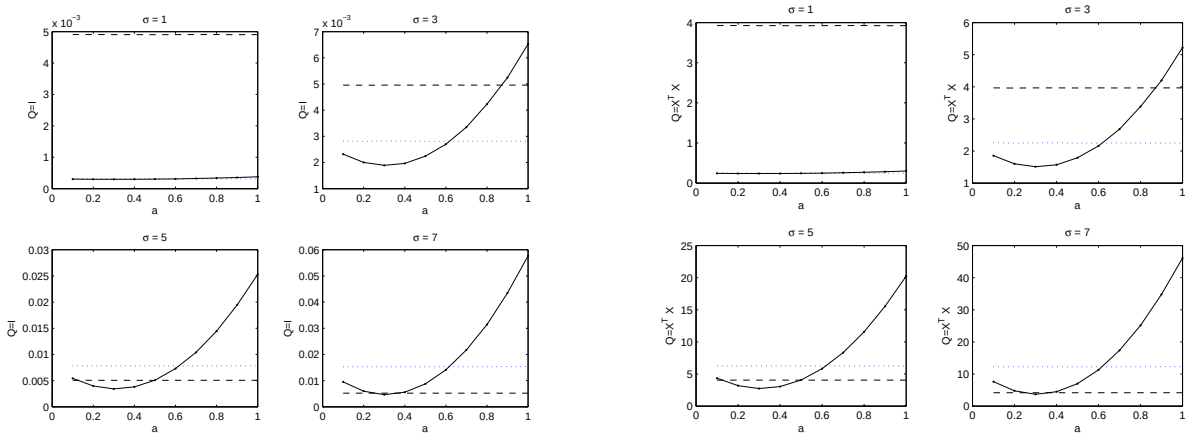

Figure 4: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are constant and equal to two. Other parameters are set to $L = 1$, $K_1 = 1$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

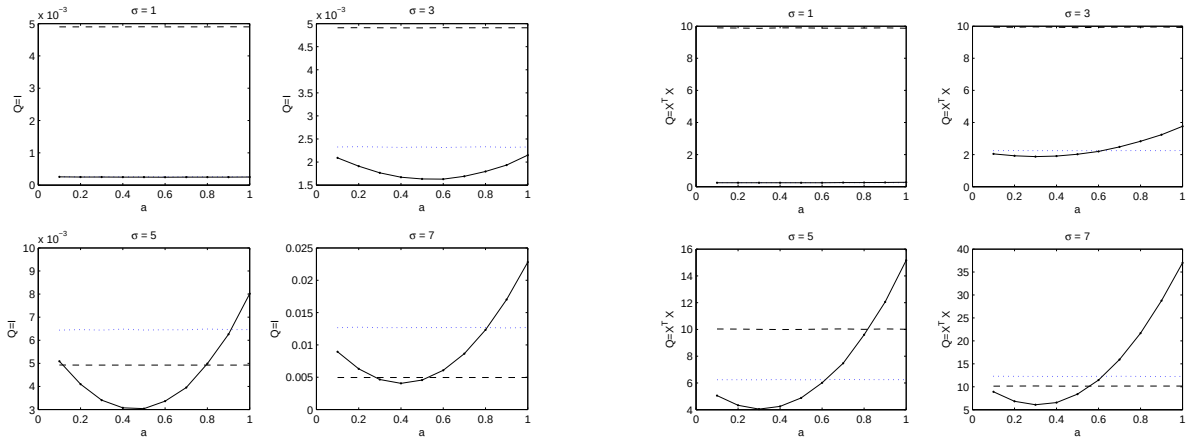

Figure 5: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are linearly decreasing from 5 to 1. Other parameters are set to $L = 1$, $K_1 = 1$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

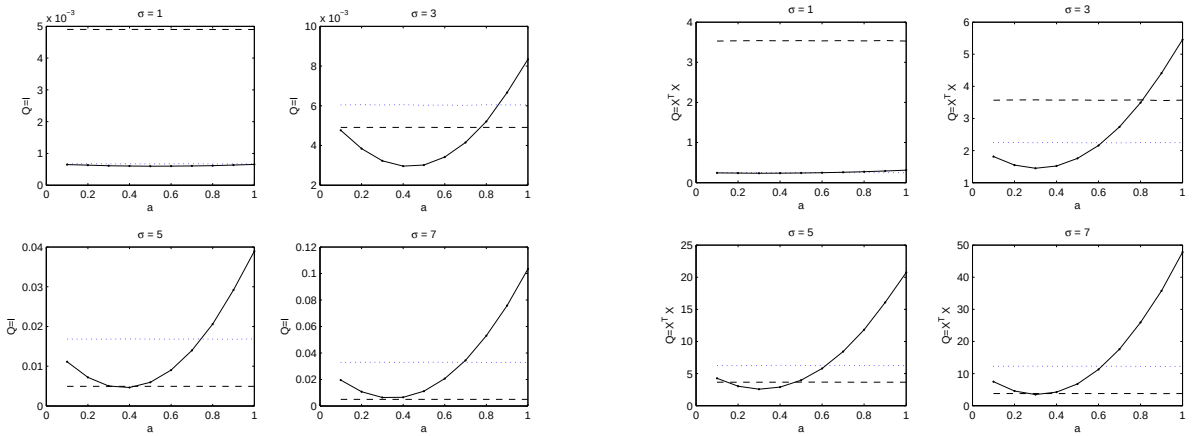

Figure 6: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $L = 1$, $K_1 = 1$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

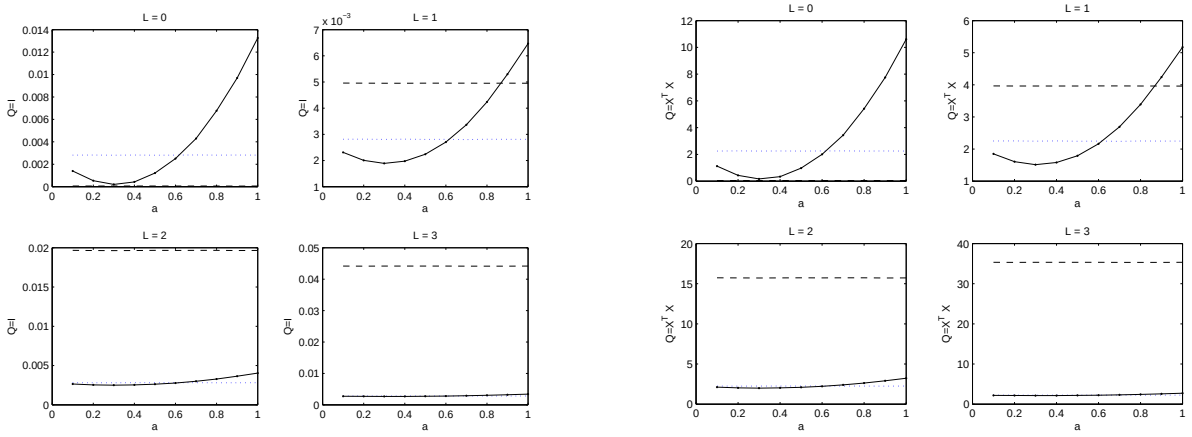

Figure 7: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are constant and equal to two. Other parameters are set to $K_1 = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

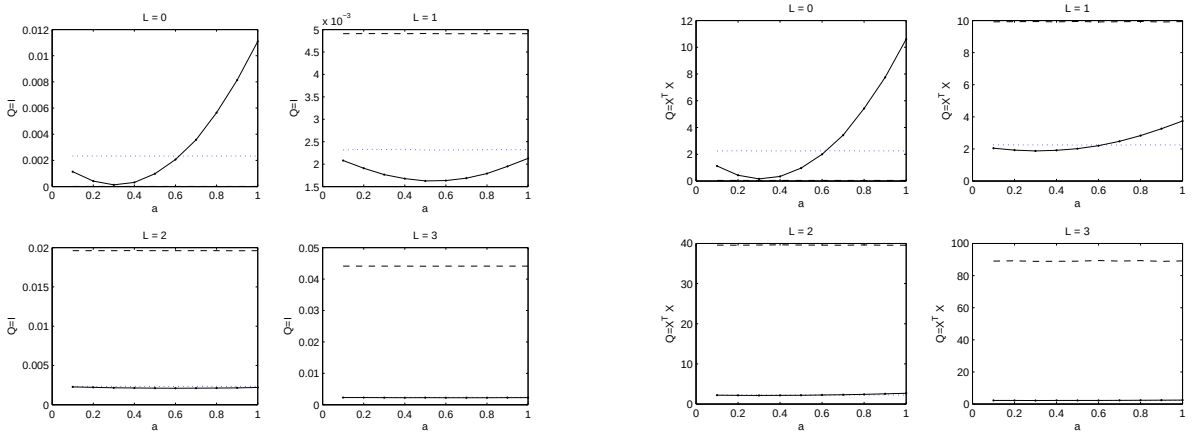

Figure 8: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are linearly decreasing from 5 to 1. Other parameters are set to $K_1 = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

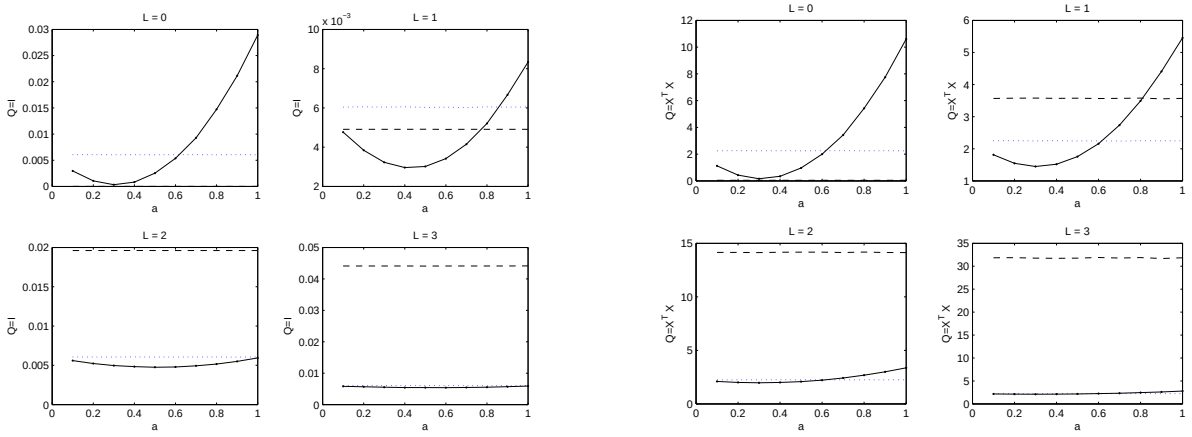

Figure 9: Risk$(\beta(y,X)) = \mathbb{E}[(\beta(y,X)-\beta)'Q(\beta(y,X)-\beta)]$ as function of a. Left panel: $Q = I$ (MSE), right panel: $Q = X'X$. The data-generating singular values are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $K_1 = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

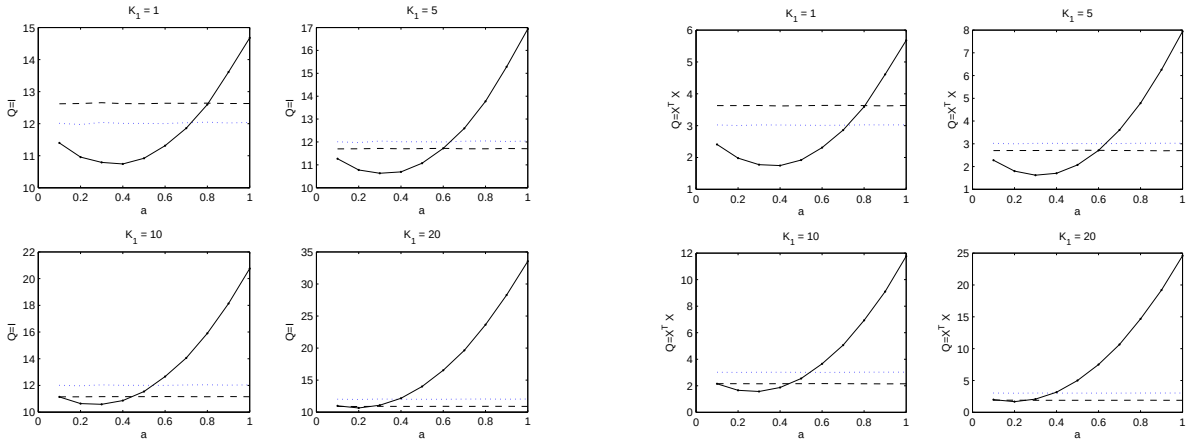

Figure 10: Left panel MSFE$(\beta(y,\, X)) = \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk$(\beta,\, \beta(y, X),\, X'X) = \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The data-generating eigenvalues are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $L = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

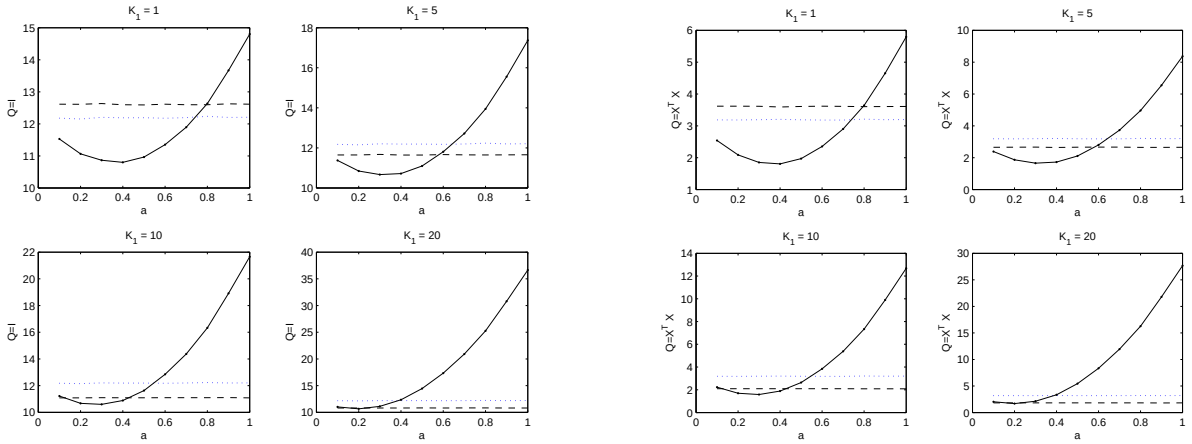

Figure 11: Left panel MSFE$(\beta(y,\, X)) = \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk$(\beta,\, \beta(y, X),\, X'X) = \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The columns of the regressor matrix X are AR(1) processes. Other parameters are set to $L = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

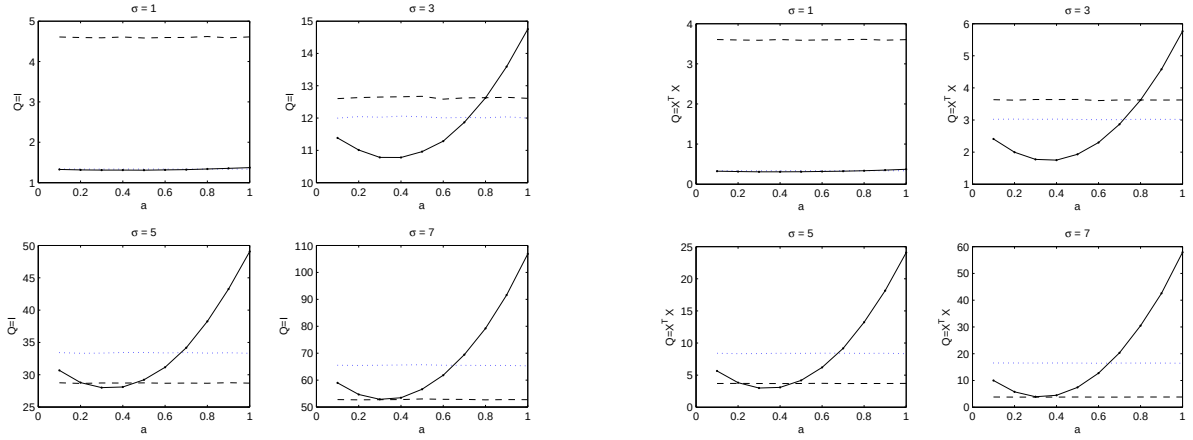

Figure 12: Left panel MSFE($\beta(y,\, X)) = \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk(β, $\beta(y, X)$, $X'X$) $= \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The data-generating eigenvalues are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $K_1 = 1$, $L = 1$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

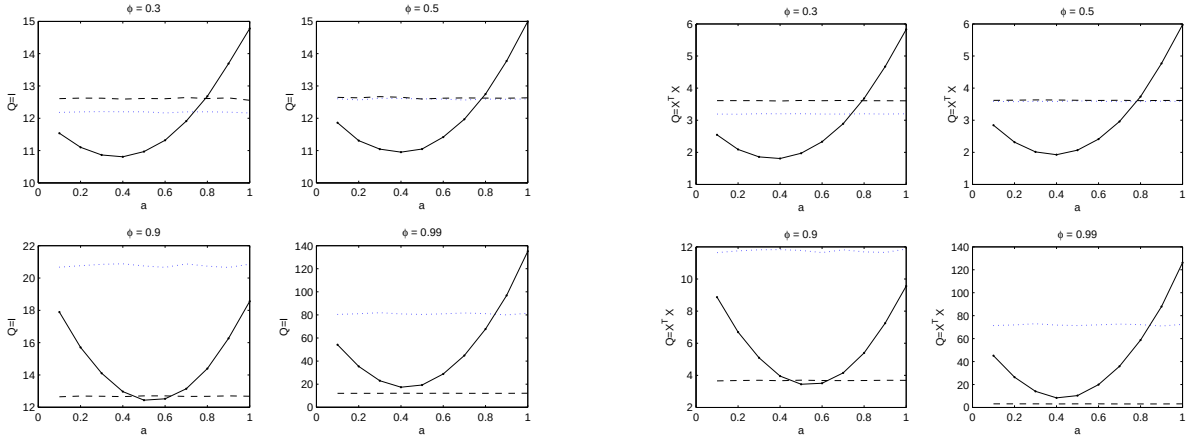

Figure 13: Left panel MSFE($\beta(y,\, X)) = \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk(β, $\beta(y, X)$, $X'X$) $= \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The columns of the regressor matrix X are AR(1) processes. Other parameters are set to $K_1 = 1$, $L = 1$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

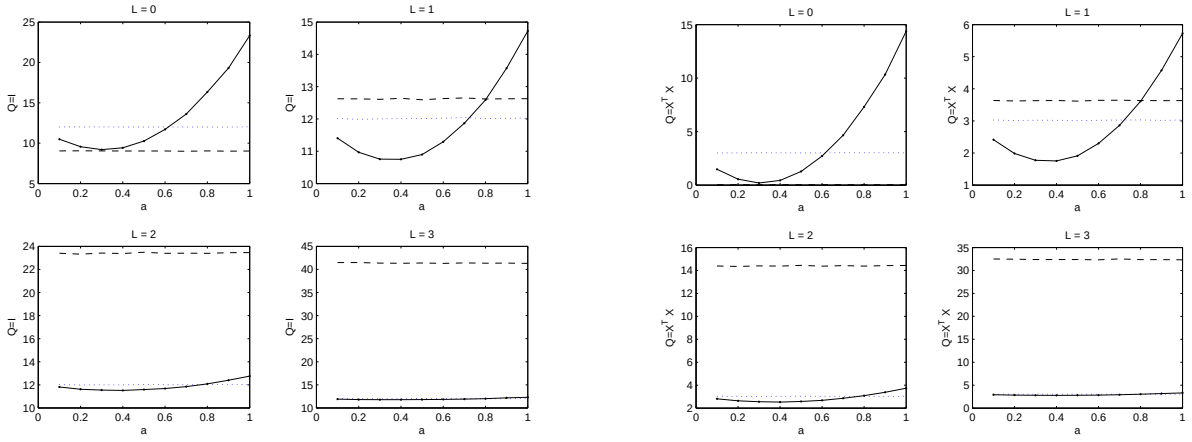

Figure 14: Left panel MSFE($\beta(y,\, X)$) $= \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk(β, $\beta(y, X)$, $X'X$) $= \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The data-generating eigenvalues are exponentially decreasing from 5 to 1 at a rate of 0.10. Other parameters are set to $K_1 = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.

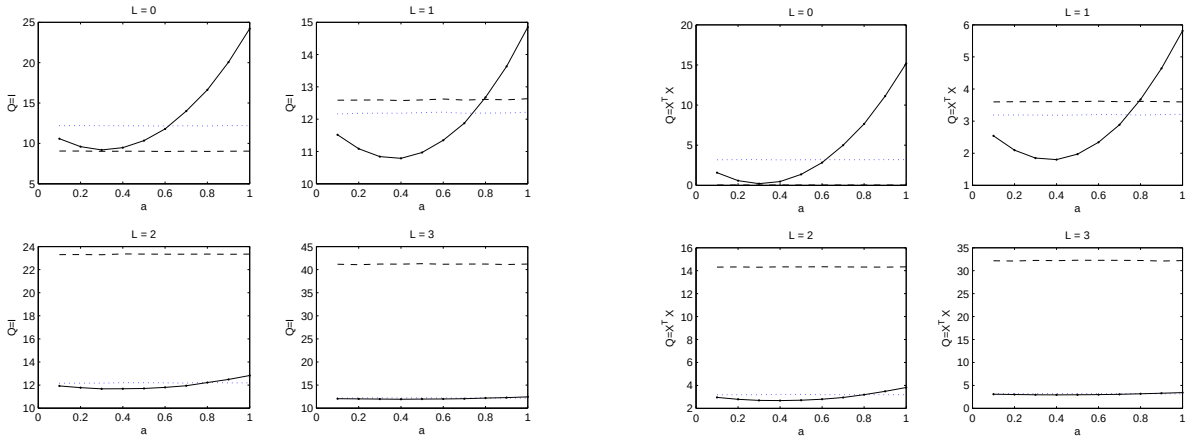

Figure 15: Left panel MSFE($\beta(y,\, X)$) $= \mathbb{E}[(\hat{y} - y)'(\hat{y} - y)]$, where $\hat{y}$ is a T_2-vector of forecasts of y, as function of a. Right panel: Risk(β, $\beta(y, X)$, $X'X$) $= \mathbb{E}[(\beta(y, X) - \beta)'X'X(\beta(y, X) - \beta)]$ for the T_2-period forecast sample. The columns of the regressor matrix X are AR(1) processes. Other parameters are set to $K_1 = 1$, $\sigma = 3$. The connected dots line shows the performance of the Stein-like estimator. For comparison, the performance of the standard OLS estimator is shown in dots. The performance of the principal components estimator is plotted with dashes.