A ridge to homogeneity for linear models

Stanislav Anatolyev*
CERGE-EI, Czech Republic and New Economic School, Russia

May 2020

Abstract

In some heavily parameterized models, one may benefit from shifting some of parameters towards a common target. We consider L^2 shrinkage towards an equal parameter value that balances between unrestricted estimation (i.e., allowing full heterogeneity) and estimation under equality restriction (i.e. imposing full homogeneity). The penalty parameter of such ridge regression estimator is tuned using leave-one-out cross-validation. The reduction in predictive mean squared error tends to increase with the dimensionality of the parameter set. We illustrate the benefit of such shrinkage with a few stylized examples. We also work out an example of a heterogeneous panel model, including estimation on real data.

Keywords: shrinkage, homogeneity restrictions, ridge regression, predictive mean squared error, cross-validation, heterogeneous panel data

*Address: Stanislav Anatolyev, CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: stanislav.anatolyev@cerge-ei.cz. This research was supported by the grants 17-26535S and 20-28055S from the Czech Science Foundation and the Access Industries professorship from the New Economic School. I thanks the Editor and two anonymous referees for useful suggestions, and also Wessel van Wieringen, Lukáš Lafférs and Daniel Henderson for valuable comments. This research was presented at the Workshop in Model Selection, Regularization, and Inference in Vienna, the 12th International Conference on Computational and Financial Econometrics in Pisa, the 5th Conference of Deutsche Arbeitsgemeinschaft Statistik in Munich, and the Czech Economic Society and Slovak Economic Association Meeting in Brno.

1 Introduction

The ridge regression estimator was originally devised as an anti-collinearity tool (Hoerl, 1959). It also has a property of trading off variance for non-zero bias and achieving lower estimation mean squared error (MSE) than that of least squares (LS) estimation. The ridge regression estimator results from a penalized LS problem with L^2 penalization. Although the typical aim of ridge regression and L^2 shrinkage in general is to produce an estimator with a lower MSE (Hoerl and Kennard, 1970; Theobald, 1974), this ability carries over to the predictive MSE as well. For more on ridge regression, see Gruber (2010) and van Wieringen (2019).

Some models are heavily parameterized, with a subset of the parameter set representing the same feature which may (or may not) be common within this subset. In that case, researchers sometimes exploit either totally unrestricted or, conversely, tightly constrained versions of the model. One important practical example is a heterogeneous coefficients panel data model, for which, despite the fact that homogeneity of coefficients across units is often rejected, the homogeneous specification is most often adopted in practice, even though a variety of shrinkage methods for heterogeneous panel data models have been developed (Baltagi, Bresson and Pirotte, 2008). Another class of applications contains various nonlinear multivariate volatility models in finance, like the BEKK models (Engle and Kroner, 1995), autoregressive Wishart models (Golosnoy, Gribisch, and Liesenfeld, 2012) and spatial multivariate GARCH (Caporin and Paruolo, 2015), where homogeneity restrictions are sometimes imposed to reduce dimensionality.

Usually, the ridge penalization is performed towards a zero target (Hoerl and Kennard, 1970); a rarer case is a constant or random non-zero target (Swindel, 1976). In this paper, we analyze choosing, as a target, the homogeneity restriction when it is relevant. The idea is to balance between unrestricted estimation (i.e., allowing full heterogeneity) and estimation under the commonality restriction (i.e. imposing full homogeneity). The L^2 penalized mean squared error criterion results in a ridge regression type estimator that has a closed form in linear models. In other contexts such as volatility models references above, one may potentially use another L^2 penalized loss function. The penalty parameter of the ridge regression estimator is tuned to minimize the out-of-sample mean squared error, which in practice can be implemented using leave-one-out cross-validation. Apart from the cross-validation stage, the proposed estimator is one-step, in contrast to an alternative shrinkage scheme where the shrinkage estimator is a weighted average between a fully unrestricted and fully restricted estimators (Hansen, 2016).

We closely consider three prototype models, one one-dimensional and two two-dimensional ones, with different ratios for the number of targets to the number of parameters that are subject to shrinkage and to the total number of observations. Denote the cardinality of the regressor vector by P, and the cross-sectional dimension by n (in the asymptotic analysis, $n \to \infty$ while P stays fixed or may slowly grow). In a simple one-dimensional prototype model, where we derive the analytical expression for the predictive MSE and the optimal value for the penalty parameter, this ratio equals $1 : P : n$. In the more complex prototype models where we derive the ridge estimators in a closed form but obtain predictive MSE by simulation of the feasible procedure, these ratios are $1 : P^2 : Pn$ and $P : P^2 : Pn$. As our prototype models indicate, there is a reduction in the predictive mean squared error which tends to increase with dimensionality of the parameter vector that is subject to shrinkage.

We also specialize to a setup of a linear heterogenous panel data model and derive the ridge estimator in a closed form. The aforementioned ratios in this case are $k : kN : NT$, where k is a number of covariates, N is a number of cross-sectional units, and T is time-series dimensionality (in the asymptotic analysis, k stays fixed, $T \to \infty$ while N stays fixed or may slowly grow). The leave-one-out cross-validation is performed along the time dimension. We illustrate this application with an empirical example of a public capital productivity model of Baltagi and Pinnoi (1995).

The paper is organized as follows. In Section 2, to fix ideas, we describe the optimization problem that leads to a ridge regression towards homogeneity. In Section 3, we work out a one-dimensional prototype model for which we derive an analytical solution and analyze asymptotic out-of-sample MSE. In Section 4, we describe two two-dimensional prototype models, derive the ridge solutions for them, and compare MSE properties of the heterogeneous, homogeneous and ridge estimators. In Section 5, we apply the idea of ridge towards homogeneity to the heterogeneous panel data setup. Section 6 concludes.

A word on notation is due. By $\text{diag}\{A_i\}_{i=1}^m$ we denote a block-diagonal matrix containing square matrices A_i, $i = 1, ..., m$, on the main diagonal, and by $\|a_i\|_{i=1}^m$ we denote a vector containing vectors a_i, $i = 1, ..., m$, stacked upon each other. Vector ι_m is an $m \times 1$ vector of ones, $J_m = \iota_m \iota_m'$ is an $m \times m$ matrix of ones, and I_m is an identity matrix of size m. We also introduce a special matrix

$$\Xi_m = I_m - \frac{J_m}{m}. \tag{1}$$

Note that Ξ_m is symmetric and idempotent.

2 Shrinkage towards homogeneity

Let t index observations from 1 to n. Let B_1, B_2, ..., B_K be non-overlapping subsets of the parameter vector/matrix B, such that each subvector/submatrix B_k contains parameters that are subject to shrinkage towards a common value, say $\bar{\beta}_k$. Let $B_- = B \backslash (B_1 \cup B_2 \cup ... \cup B_K)$. Let $e_t(B)$ be the vector of regression residuals for observation $t = 1, ..., n$ when the values of parameters are fixed at B, i.e., for example, $e_t(B) = Y_t - BX_t$ as in models of Section 4. The L^2 penalized mean squared error criterion is

$$\min_{B_1,...,B_K,B_-} \left\{ PLS(\lambda_\beta) = \frac{1}{2} \sum_{t=1}^{n} e_t(B_1, ..., B_K, B_-)' e_t(B_1, ..., B_K, B_-) + \frac{\lambda_\beta}{2} \sum_{k=1}^{K} \sum_{\beta_{jk} \in B_k} \left(\beta_{jk} - \bar{\beta}_k\right)^2 \right\}, \tag{2}$$

where $\lambda_\beta > 0$ is a degree of penalization parameter, and

$$\bar{\beta}_k = \frac{1}{|B_k|} \sum_{\beta_{jk} \in B_k} \beta_{jk} \tag{3}$$

is the common target within B_k.

The first order conditions to (2) are written as follows. For each $\beta_{jk} \in B_k$, $k = 1, ..., K$, that is subject to shrinkage within the parameter subset B_k,

$$0 = \sum_{t=1}^{n} e_t(B_1, ..., B_K, B_-)' \frac{\partial e_t(B_1, ..., B_K, B_-)}{\partial \beta_{jk}} + \lambda_\beta \left(\left(\beta_{jk} - \bar{\beta}_k\right) - \frac{1}{|B_k|} \sum_{\beta_{jk} \in B_k} \left(\beta_{jk} - \bar{\beta}_k\right) \right),$$

or

$$0 = \sum_{t=1}^{n} e_t(B_1, ..., B_K, B_-)' \frac{\partial e_t(B_1, ..., B_K, B_-)}{\partial \beta_{jk}} + \lambda_\beta \left(\beta_{jk} - \bar{\beta}_k\right). \tag{4}$$

For each $\beta_j \in B_-$ that is not subject to shrinkage, we have the usual least squared equation

$$0 = \sum_{t=1}^{n} e_t(B_1, ..., B_K, B_-)' \frac{\partial e_t(B_1, ..., B_K, B_-)}{\partial \beta_j}. \tag{5}$$

Let us denote the solution to (4)–(5) by $\hat{B}_{\text{ridge}}$. Of course, the case $\lambda_\beta = 0$ corresponds to the unconstrained (heterogeneous) solution, say $\hat{B}_{\text{het}}$, while the case $\lambda_\beta \to \infty$ results in all $\beta_{jk} \in B_k$, $k = 1, ..., K$, being equal, i.e. to the fully constrained (homogeneous) solution, say $\hat{\beta}_{\text{hom}}$.

When $e_t(B_1, ..., B_K, B_-)$ is linear in elements of all B_1, B_2, ..., B_K, B_-, the FOC are linear in all these elements, and there is a closed form solution. We illustrate this with three models in the next two Sections.

Regarding the asymptotic inference, as long as $|B|$ stays fixed asymptotically as $n \to \infty$, the asymptotic properties of $\hat{B}_{\text{ridge}}$ for any fixed λ_β are the same as those of $\hat{B}_{\text{het}}$. We conjecture that $|B|$ may be allowed to slowly increase as $n \to \infty$, without drastic changes in asymptotic conclusions. For a linear model, the finite sample conditional variance is easily available for any fixed value of the penalty parameter λ_β.

3 One-dimensional prototype model

3.1 Setup

Consider a regression equation with P regressors and similar coefficients:

$$y_t = x_t'\beta + e_t, \quad t = 1, ..., n, \tag{6}$$

where for simplicity e_t is IID across t with mean zero and variance σ^2, the $P \times 1$ vectors x_t are IID $N(0, I_P)$. The coefficient vector β contains elements β_j that are close to each other so that the bias-variance tradeoff makes the homogeneous and heterogeneous estimates competitive. For concreteness, we adopt a 'random design' setup and assume that $\beta_j \sim$ IID $N(\beta_0, \omega_\beta)$ for some concentration point β_0 and degree of heterogeneity ω_β. The researcher does not know these parameters but has a prior belief about similarity of the elements of β, which is a reason to impose the homogeneity restriction for the sake of model parsimony. The ridge machinery can be used to further exploit the bias-variance tradeoff. In this model, the ratio for the number of targets to the number of parameters that are subject to shrinkage and to the total number of observations equals $1 : P : n$.

3.2 Estimation

If elements of β are treated as distinct parameters (i.e. allowing full heterogeneity) and estimated separately, we use least squares estimates

$$\hat{\beta}_{\text{het}} = \left(\sum_{t=1}^{n} x_t x_t' \right)^{-1} \sum_{t=1}^{n} x_t y_t. \tag{7}$$

If one makes all the elements of β equal (i.e. imposing full homogeneity), then the homogeneous solution is

$$\hat{\beta}_{\text{hom}} = \left(\sum_{t=1}^{n} (x_t'\iota_P)^2 \right)^{-1} \sum_{t=1}^{n} (x_t'\iota_P)(y_t\iota_P). \tag{8}$$

Suppose we use the following penalized estimation criterion:

$$\min_{\beta}\left\{PLS\left(\lambda_{\beta}\right)=\frac{1}{2}\sum_{t=1}^{n}\left(y_{t}-x_{t}'\beta\right)^{2}+\frac{\lambda_{\beta}}{2}\sum_{j=1}^{P}\left(\beta_{j}-\bar{\beta}\right)^{2}\right\} \tag{9}$$

where

$$\bar{\beta}=\frac{1}{P}\sum_{j=1}^{P}\beta_{j}. \tag{10}$$

Solving (9) leads to the ridge regression estimator[1]

$$\hat{\beta}_{\text{ridge}}=\left(\sum_{t=1}^{n}x_{t}x_{t}'+\lambda_{\beta}\Xi_{P}\right)^{-1}\sum_{t=1}^{n}x_{t}y_{t}. \tag{11}$$

3.3 Asymptotic predictive MSE

Next, we derive an asymptotic predictive (out-of-sample) mean squared error of a forecast generated by the ridge estimator, in order to see how it varies with the penalty parameter λ_{β}. The predictive MSE is defined as

$$MSE_{\text{ridge}}=E\left[\left(y_{t}^{*}-x_{t}^{*\prime}\hat{\beta}_{\text{ridge}}\right)^{2}\right], \tag{12}$$

where the pair (y_t^*, x_t^*) is drawn from the population of (y_t, x_t) independently from the given sample. The asymptotic analog $AMSE_{\text{ridge}}$ omits higher-order asymptotic terms in the expansion of MSE_{ridge} (12) in powers of $1/n$.

Proposition 1. The asymptotic (as $n \to \infty$ and P stays fixed) predictive MSE to order $O\left(1/n^{2}\right)$ of the forecast generated by $\hat{\beta}_{\text{ridge}}$ is

$$AMSE_{\text{ridge}}=\sigma^{2}\left(1+\frac{P}{n}+\frac{P\left(P+1\right)}{n^{2}}\right)+\left(\omega_{\beta}\lambda_{\beta}^{2}-2\sigma^{2}\lambda_{\beta}\right)\frac{P-1}{n^{2}}. \tag{13}$$

As follows from the proof of Proposition 1 (see Appendix), the first of five terms in the expression (13), σ^2, is the prediction error variance; the second component $\sigma^2 P/n$ is the estimation noise coming from the error term; the third term $\sigma^2 P\left(P+1\right)/n^2$ is the estimation noise coming from approximating population moments by sample moments; the fourth component $\omega_{\beta}\lambda_{\beta}^{2}\left(P-1\right)/n^{2}$ is an increase in estimation bias because of coefficient

[1]Note that the penalty term, apart from a multiplicative constant, can be rewritten as

$$\left(||\beta_{j}||_{j=1}^{P}-\bar{\beta}\right)'\left(||\beta_{j}||_{j=1}^{P}-\bar{\beta}\right)=\left(\Xi_{P}||\beta_{j}||_{j=1}^{P}\right)'\left(\Xi_{P}||\beta_{j}||_{j=1}^{P}\right)=\left(||\beta_{j}||_{j=1}^{P}\right)'\Xi_{P}\left(||\beta_{j}||_{j=1}^{P}\right).$$

Hence, this estimator can be interpreted as a generalized ridge estimator with weight matrix Ξ_P. I thank Wessel van Wieringen for this observation.

heterogeneity; finally, the fifth term $2\sigma^2\lambda_\beta (P-1)/n^2$ is a reduction in MSE because of a decrease in the estimation variance. While the first three terms are indispensable, the last two terms represent the bias-variance tradeoff in the MSE. Note that this tradeoff is of order $O(1/n^2)$, unless P changes with n, and quickly falls with sample size, as is usual with ridge-shrinkage. Note also that the asymptotic MSE (13) does not depend on the parameter concentration point β_0.

Remark. Normality imposed on the distribution of x and β_j is not critical for the expression (13) to hold as long as the second-order moments exist.

Remark. The expression (13) also obtains if asymptotically, as $n \to \infty$, P also grows sufficiently slowly; in particular, if $P = o(\sqrt{n})$.

Now observe that $\partial AMSE_{\text{ridge}}/\partial\lambda_\beta|_{\lambda_\beta=0} < 0$, hence, at least for a range of $\lambda_\beta > 0$ when the degree of heterogeneity ω_β is small relative to the error variance σ^2, we have $AMSE_{\text{ridge}} < AMSE_{\text{het}}$. The optimal λ_β can be obtained from

$$0 = \frac{\partial AMSE_{\text{ridge}}}{\partial\lambda_\beta} = \left(2\omega_\beta\lambda_\beta - 2\sigma^2\right)\frac{P-1}{n^2}, \tag{14}$$

provided that $\omega_\beta \neq 0$, which leads to

$$\lambda_\beta^{\text{opt}} = \frac{\sigma^2}{\omega_\beta}. \tag{15}$$

That is, more shrinkage is required if there is more error noise and less coefficient heterogeneity. The minimal $AMSE_{\text{ridge}}$ is

$$AMSE_{\text{ridge}}^{\text{opt}} = \sigma^2\left(1 + \frac{P}{n} + \frac{P(P+1)}{n^2}\right) - \frac{\sigma^4}{\omega_\beta}\frac{P-1}{n^2}. \tag{16}$$

Clearly, as $\lambda_\beta \to \infty$, the increase in MSE will dominate the reduction in MSE, ridge will not be beneficial. In fact, this will occur when λ_β reaches or exceeds $2\lambda_\beta^{\text{opt}}$. At the other end, setting $\lambda_\beta = 0$ is equivalent to imposing no constraints, which leads to the heterogeneous solution

$$AMSE_{\text{het}} = \sigma^2\left(1 + \frac{P}{n} + \frac{P(P+1)}{n^2}\right). \tag{17}$$

The maximal benefit in the MSE of the ridge over the unconstrained estimation is

$$AMSE_{\text{het}} - AMSE_{\text{ridge}}^{\text{opt}} = \frac{\sigma^4}{\omega_\beta}\frac{P-1}{n^2}, \tag{18}$$

which increases in the subject-to-shrinkage parameter dimensionality P.

3.4 Feasible ridge estimator

In practice, the optimal value of the penalty parameter λ_β is unknown. We suggest that the feasible version, which we call 'CV-ridge', be based on leave-one-out cross-validation: at the i^{th} step, $i = 1, ..., n$, the vector β is estimated based on all observations except i^{th}, and the i^{th} contribution to MSE is evaluated using the i^{th} observation on regressor and outcome variables.

3.5 Performance of feasible estimators

At each simulation run, we evaluate the four estimators' MSE using 10,000 extra pseudo-observations generated according to the same DGP. The sample size is $n = 100$; the number of simulations is 1000. We set the point of parameter concentration to $\beta_0 = 1$. We make experiments with several values of degree of heterogeneity $\omega_\beta \in \{0.025; 0.05; 0.10; 0, 15; 0.20\}$. To get a feel of the degree of dispersion in β_j's, we show in Table 1 fragments of β corresponding to each value of ω_β. For the feasible ridge estimator, the penalty parameter λ_β is selected from a grid of values from 0 to 1,000 with a step of 10.

Table 1: Fragments of β for different degrees of heterogeneity ω_β.

$\omega_\beta = 0.025$	$\omega_\beta = 0.05$	$\omega_\beta = 0.10$	$\omega_\beta = 0.15$	$\omega_\beta = 0.20$
$\begin{bmatrix} 1.015 \\ 0.968 \\ 1.024 \end{bmatrix}$	$\begin{bmatrix} 0.967 \\ 1.056 \\ 1.027 \end{bmatrix}$	$\begin{bmatrix} 0.920 \\ 1.061 \\ 0.873 \end{bmatrix}$	$\begin{bmatrix} 0.902 \\ 1.097 \\ 0.968 \end{bmatrix}$	$\begin{bmatrix} 1.478 \\ 0.715 \\ 0.869 \end{bmatrix}$

The graphs on Figure 1 present the average (across simulations) predictive MSE. The top panel shows how the MSE varies with the degree of heterogeneity ω_β when P is fixed at 20, while the bottom panel shows its dependence on the parameter dimensionality P when ω_β is fixed at 0.10.

As the degree of heterogeneity goes up, the MSE from the heterogeneous estimates (orange lines) remains fully flat, while the MSE from the homogeneous estimates (gray lines) quickly increases and from some point exceeds the former. The optimal ridge estimates (yellow lines) and CV ridge estimates (blue lines) dominate in terms of MSE, though their performance approaches that of heterogeneous estimates as the degree of heterogeneity increases. Note that the discrepancy between the MSE from the infeasible optimal ridge and

feasible CV ridge estimation is negligible compared to the differences between these and MSE from homo- and heterogeneous estimates.

Further, as the parameter dimensionality increases, the MSE from all estimates rises, from heterogeneous estimates more so, while from the ridge estimators in the least degree. Thus, the benefit from using the ridge machinery tends to reveal itself in a higher degree for higher dimensional setups.

It is also instructive to check on the values of the penalty parameter $\lambda_\beta^{\text{CV}}$ in the feasible CV-ridge procedure compared to the asymptotically optimal. The median (across simulations) values are presented in Table 2 for the same values of degree of heterogeneity ω_β and $P = 20$. One can see that the selected via cross-validation penalty parameter is not far (in median terms) from the ideal optimal value, though it may be either larger or smaller than that. The evidence presented in Figure 2 suggests that the discrepancy is not critical for the predictive performance.

Table 2: Comparison of asymptotically optimal and feasible values of penalty parameter λ_β for different degrees of heterogeneity ω_β.

ω_β	0.025	0.05	0.10	0.15	0.20
$\lambda_\beta^{\text{opt}}$	1600	400	100	44	25
$\lambda_\beta^{\text{CV}}$	1000	470	100	40	30

4 Two-dimensional prototype models

In this Section, we consider more complex examples and compare MSE of different estimators by simulations. These are characterized by two-dimensional design, i.e. a number of parameters subject to shrinkage is P^2, with a number of targets being equal to 1 or P. Again, in addition to the heterogeneous, homogeneous and non-feasible optimal ridge estimators, we also consider the feasible version of the optimal ridge estimator, with the optimal smoothing parameter obtained via leave-one-out cross-validation.

4.1 Setup

In both models, there is a system of P non-homogeneous equations

$$Y_t = BX_t + e_t, \quad t = 1, ..., n. \tag{19}$$

where for simplicity e_t is IID across t with mean zero and variance matrix $\sigma^2 I_P$, with $\sigma^2 = 1$. The coefficient matrix B is

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1P} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{P1} & \beta_{P2} & \cdots & \beta_{PP} \end{bmatrix}. \tag{20}$$

We generate B such that β_{jk} are concentrated around the same concentration point β_0: $\beta_{jk} \sim IIDN(\beta_0, \omega_\beta)$, where, again, ω_β is degree of heterogeneity. Of course, the researcher does not know these parameters but has a prior belief that the elements of B are close to each other, which is the reason to impose, fully or partially, the homogeneity restriction and exploit the ridge trade-off. For simplicity, we generate X_t as independent standard normals independently of B and all e_t.

In these models, the ratio for the number of targets to the number of parameters that are subject to shrinkage and to the total number of observations equals $1 : P^2 : Pn$ in the 'fully homogeneous target' model, and $P : P^2 : Pn$ in the 'row-wise homogeneous target' model.

4.2 Fully homogeneous target

The elements of B may be estimated without restrictions, imposing the homogeneity restrictions, or balance between the two by penalizing deviations from the common fully homogeneous target. If elements of B are treated as distinct parameters (i.e. allowing full heterogeneity) and estimated separately, we use least squares estimates

$$\hat{B}_{\text{het}} = \sum_{t=1}^{n} Y_t X_t' \left(\sum_{t=1}^{n} X_t X_t' \right)^{-1}. \tag{21}$$

If all elements of B are equated (i.e. under full homogeneity restriction), then

$$B = \beta J_P. \tag{22}$$

The least squares estimation problem is then

$$\min_{\beta} \frac{1}{2} \sum_{t=1}^{n} e_t(\beta)' e_t(\beta), \tag{23}$$

where $e_t(\beta) = Y_t - \beta J_P X_t$. The solution to (23) is $\hat{B}_{\text{hom}} = \hat{\beta}_{\text{hom}} J_P$, where

$$\hat{\beta}_{\text{hom}} = \left(P \sum_{t=1}^{n} X_t' J_P X_t \right)^{-1} \sum_{t=1}^{n} X_t' J_P Y_t, \tag{24}$$

because $J_P^2 = PJ_P$.

Suppose we use the following penalized estimation criterion:

$$\min_B \left\{ PLS\left(\lambda_\beta\right) = \frac{1}{2}\sum_{t=1}^{n} e_t\left(B\right)' e_t\left(B\right) + \frac{\lambda_\beta}{2}\sum_{j,k=1}^{P}\left(\beta_{jk} - \bar{\beta}\right)^2 \right\}, \tag{25}$$

where

$$\bar{\beta} = \frac{1}{P^2}\sum_{j,k=1}^{P}\beta_{jk}. \tag{26}$$

The solution to (25) is ridge-type regression estimator:

$$vec\left(\hat{B}'_{\text{ridge}}\right) = \left(I_P \otimes X'X + \lambda_\beta \Xi_{P^2}\right)^{-1} vec\left(X'Y\right). \tag{27}$$

The smoothing parameter λ_β is set to minimize the out-of-sample mean squared prediction error criterion. The feasible version which we call 'CV' is based on leave-one-out cross-validation: at the i^{th} step, $i = 1, ..., n$, the matrix B is estimated based on all observations except i^{th}, and the i^{th} contribution to MSE is evaluated using the i^{th} observation on regressor and outcome variables.

4.3 Row-wise homogeneous target

In this setup, the researcher instead has a belief that the elements of each row of the matrix B are close to each other but may not be close across the rows. In this case, the researcher may want to impose the homogeneity restriction partially, only within each row. Then, the target is a P-vector of row-specific values.

The full heterogeneity estimate $\hat{B}_{\text{het}}$ is the same as before. Under row-wise homogeneity restriction, all elements of each row of B are equated, so that

$$B = \beta \iota'_P, \tag{28}$$

where now β is a $P \times 1$ column vector. The least squares estimation problem is

$$\min_\beta \frac{1}{2}\sum_{t=1}^{n} e_t\left(\beta\right)' e_t\left(\beta\right), \tag{29}$$

where $e_t\left(\beta\right) = Y_t - \beta \iota'_P X_t$. The solution to (29) is $\hat{B}_{\text{hom}} = \hat{\beta}_{\text{hom}} \iota'_P$, where

$$\hat{\beta}_{\text{hom}} = \left(\sum_{t=1}^{n}\left(\iota'_P X_t\right)^2\right)^{-1}\sum_{t=1}^{n}\left(\iota'_P X_t\right) Y_t. \tag{30}$$

Now, suppose we use the following penalized estimation criterion:

$$\min_{B}\left\{PLS\left(\lambda_{\beta}\right)=\frac{1}{2}\sum_{t=1}^{n}e_{t}\left(B\right)'e_{t}\left(B\right)+\frac{\lambda_{\beta}}{2}\sum_{j=1}^{P}\sum_{k=1}^{P}\left(\beta_{jk}-\bar{\beta}_{j}\right)^{2}\right\}, \tag{31}$$

where

$$\bar{\beta}_{j}=\frac{1}{P}\sum_{k=1}^{P}\beta_{jk}. \tag{32}$$

The solution to (31) is ridge-type regression estimator:

$$vec\left(\hat{B}'_{\text{ridge}}\right)=\left(I_{P}\otimes\left(X'X+\lambda_{\beta}\Xi_{P}\right)^{-1}\right)vec\left(X'Y\right). \tag{33}$$

Again, the smoothing parameter λ_{β} is set to minimize the out-of-sample mean squared prediction error criterion. The feasible version which we call 'CV' is based on leave-one-out cross-validation: at the i^{th} step, $i=1,...,n$, the matrix B is estimated based on all observations except i^{th}, and the i^{th} contribution to MSE is evaluated using the i^{th} observation on regressor and outcome variables.

4.4 Comparison of predictive MSE

At each simulation run, we evaluate the four estimators' MSE using 10,000 extra pseudo-observations generated according to the same DGP. The sample size is $n=200$ for the fully homogeneous target example, and $n=100$ for the row-wise homogeneous target example. The number of simulations is 100. We set the point of concentration of parameters at $\beta_{0}=1$. We make experiments with several values of degree of heterogeneity $\omega_{\beta}\in\{0.025;0.05;0.10;0,15;0.20\}$. To get a feel of the degree of dispersion in B, we show in Table 3 fragments of B corresponding to each value of ω_{β}.

Table 3: Fragments of B for different degrees of heterogeneity ω_{β}.

$\omega_{\beta}=0.025$	$\omega_{\beta}=0.05$	$\omega_{\beta}=0.10$	$\omega_{\beta}=0.15$	$\omega_{\beta}=0.20$
$\begin{bmatrix} 0.980 & 1.015 \\ 1.024 & 0.981 \end{bmatrix}$	$\begin{bmatrix} 0.960 & 1.030 \\ 0.937 & 1.048 \end{bmatrix}$	$\begin{bmatrix} 0.920 & 1.061 \\ 0.873 & 1.096 \end{bmatrix}$	$\begin{bmatrix} 0.879 & 1.091 \\ 0.809 & 1.144 \end{bmatrix}$	$\begin{bmatrix} 0.839 & 1.122 \\ 0.746 & 1.192 \end{bmatrix}$

The graphs on Figures 2 and 3 present the average (across simulations) MSE for the fully homogeneous and the row-wise homogeneous targets, respectively. The top panel shows how the MSE varies with the degree of heterogeneity ω_{β} when P is fixed at 20, while the bottom

panel shows its dependence on the parameter dimensionality P when ω_β is fixed at 0.05 (Figure 2) or 0.10 (Figure 3).

The two cases exhibit qualitatively similar tendencies. Moreover, all tendencies are similar to those for the one-dimensional prototype model. It is interesting though that for the row-wise homogeneous target this benefit is much higher than for the fully homogeneous target, even though the former target is less homogeneous, especially with large P.

5 Application: heterogeneous panel data

Consider the panel data model with cross-sectionally heterogeneous coefficients:

$$y_{i,t} = x'_{i,t}\beta_i + \mu_i + v_{i,t}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{34}$$

where the idiosyncratic components $v_{i,t}$ are IID across i and t, and, conditional on $X = (x_{1,1}, ..., x_{N,T})$, have zero mean and variance σ_v^2. Here, the $k \times 1$ slope coefficients β_i vary across the individuals. After the Within transformation, we have

$$\tilde{y}_{i,t} = \tilde{x}'_{i,t}\beta_i + \tilde{v}_{i,t}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{35}$$

where $\tilde{y}_{i,t} = y_{i,t} - \bar{y}_i$, $\tilde{x}_{i,t} = x_{i,t} - \bar{x}_i$, etc. The homogeneous estimates $\hat{\beta}_{\text{hom}}$ are given by the Within estimator, i.e. OLS on the pooled Within-transformed system. The sample mean squared error is computed as an average of squared Within residuals:

$$\widehat{MSE}_{\text{hom}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{i,t} - \tilde{x}'_{i,t}\hat{\beta}_{\text{hom}}\right)^2. \tag{36}$$

Suppose now that we shrink each vector β_i to a common row-specific vector of values β. The ridge problem is

$$\min_{(\beta_1,\beta_2,...,\beta_N)}\left\{PLS\left(\lambda_\beta\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{i,t} - \tilde{x}'_{i,t}\beta_i\right)^2 + \frac{\lambda_\beta}{2}\sum_{i=1}^{N}\left(\beta_i - \bar{\beta}\right)'\left(\beta_i - \bar{\beta}\right)\right\}, \tag{37}$$

where

$$\bar{\beta} = \frac{1}{N}\sum_{i=1}^{N}\beta_i. \tag{38}$$

The first order conditions with respect to β_i for all $i = 1, ..., N$, are

$$\sum_{t=1}^{T}\tilde{x}_{i,t}\left(\tilde{y}_{i,t} - \tilde{x}'_{i,t}\beta_i\right) = \lambda\left(\beta_i - \bar{\beta}\right). \tag{39}$$

Hence, the ridge estimators stacked upon each other are

$$\left\|\hat{\beta}_{i,\text{ridge}}\right\|_{i=1}^{N} = \Psi_{N,T}\left(\lambda_\beta\right)^{-1}\left\|\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{y}_{i,t}\right\|_{i=1}^{N}, \tag{40}$$

where

$$\Psi_{N,T}\left(\lambda_\beta\right) \equiv \text{diag}\left\{\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{x}_{i,t}'\right\}_{i=1}^{N} + \lambda_\beta \Xi_N \otimes I_k. \tag{41}$$

The sample mean squared error is computed as an average of squared ridge residuals:

$$\widehat{MSE}_{\text{ridge}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{i,t} - \tilde{x}_{i,t}'\hat{\beta}_{i,\text{ridge}}\right)^2. \tag{42}$$

The heterogeneous estimator corresponds to the above solution with $\lambda_\beta = 0$:

$$\hat{\beta}_{\text{het}} = \Psi_{N,T}\left(0\right)^{-1}\left\|\sum_{t=1}^{T}\tilde{x}_{i,t}\tilde{y}_{i,t}\right\|_{i=1}^{N}, \tag{43}$$

i.e. each $\hat{\beta}_i$ is computed as OLS from a time series regression on unit i^{th} data. The sample mean squared error is computed as an average of squared heterogeneous residuals:

$$\widehat{MSE}_{\text{het}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{i,t} - \tilde{x}_{i,t}'\hat{\beta}_{i,\text{het}}\right)^2. \tag{44}$$

A variety of shrinkage methods for heterogeneous panel data models have been developed before (Baltagi, Bresson and Pirotte, 2008). One of most straightforward is the Stein-rule estimator (Ziemer and Wetzstein, 1983)

$$\hat{\beta}_{\text{Stein}} = \left(1 - \frac{c}{F}\right)\hat{\beta}_{\text{het}} + \frac{c}{F}\iota_N \otimes \hat{\beta}_{\text{hom}}, \tag{45}$$

where

$$F = \frac{RSS_{\text{hom}} - RSS_{\text{het}}}{RSS_{\text{het}}}\frac{N(T-1-k)}{(N-1)k} \tag{46}$$

is an F-statistic for pre-testing the poolability of under normality, RSS_{hom} and RSS_{het} are residual sums of squares in the homogeneous (pooled) and heterogeneous models, respectively, and

$$c = \frac{(N-1)k - 2}{N(T-1-k)+2} \tag{47}$$

(Judge and Bock, 1978).

We use the dataset used in Baltagi and Pinnoi (1995) and one of their models for public capital productivity:

$$\log Y_{i.t} = \beta_{1,i}\log P_{i.t} + \beta_{2,i}\log K_{i.t} + \beta_{3,i}\log L_{i.t} + \beta_{4,i}U_{i.t} + \mu_i + v_{i,t}, \tag{48}$$

where in state i in year t, $Y_{i.t}$ is gross state product, $P_{i.t}$ is public capital, $K_{i.t}$ is private capital, $L_{i.t}$ is labor input, and $U_{i.t}$ is the state unemployment rate. The log-linear form of the model is due to the multiplicative form of the Cobb-Douglas production function with respect to its inputs traditional in economic theory. The data embrace 48 US states for the period 1970–1986. Thus, $k = 4$, $N = 48$, and $T = 17$.

The F-statistic for across-state homogeneity of coefficients equals 7.25 with the p-value that is essentially zero. The Stein's factor $1 - c/F$ equals 0.956, so almost all the weight in the Stein-rule estimator is given to the heterogeneous estimates.

To compute the ridge estimates, we pre-scale the Within-transformed regressors by their standard deviations so that they have unit sample variance; we adjust the estimates of coefficients accordingly. The leave-one-out cross-validation is performed along the t direction. The CV-tuned penalty parameter turns out to be $\lambda_\beta = 0.0325$. The ridge sample MSE equals 0.414×10^{-3}, while the fully heterogeneous and homogeneous sample MSE equal 0.405×10^{-3} and 1.362×10^{-3}, respectively. One can see that the ridge machinery puts the solution pretty close to the heterogeneous one. However, it still has a strong effect on coefficient estimates, much stronger than the Stein shrinkage (see Table 4).

Table 4: Coefficient point estimates for states whose names start with A.

State i	$\beta_{1,i}$	$\beta_{2,i}$	$\beta_{3,i}$	$\beta_{4,i}$	$\beta_{1,i}$	$\beta_{2,i}$	$\beta_{3,i}$	$\beta_{4,i}$
	Homogeneous				Heterogeneous			
AL	−0.026	0.292	0.768	−0.005	−1.443	0.280	1.835	0.007
AZ	−0.026	0.292	0.768	−0.005	−0.163	−0.005	1.076	−0.004
AR	−0.026	0.292	0.768	−0.005	−0.506	0.321	1.234	0.001
	Stein rule				CV-ridge			
AL	−1.380	0.280	1.788	0.007	−0.915	0.259	1.520	0.005
AZ	−0.157	0.008	1.062	−0.004	0.026	0.085	0.865	−0.007
AR	−0.484	0.320	1.213	0.001	−0.418	0.329	1.172	0.001

Table 4 reports homogeneous, heterogeneous, Stein rule, and CV-ridge estimates for three arbitrary states. Of course, the homogeneous estimates are equal across the states. The heterogeneous, Stein and ridge estimates vary a lot across the states, though the dispersions of the three are different: the standard deviations are 0.55/0.35/0.52/0.0114, 0.53/0.33/0.50/0.0109 and 0.49/0.31/0.44/0.0099, respectively. Thus, the ridge estimators

exhibit a lower degree of heterogeneity of coefficient estimates than the Stein rule.

6 Concluding remarks

We have developed an extension of the ridge regression estimator to the case where the shrinkage targets are commonality, or homogeneity, of parameter subsets. This yields a reduction in a predictive mean squared error, this reduction being positively related to a number of parameters that are subject to shrinkage. The penalty parameter is tuned by minimization of the leave-one-out sample predictive cross-validation criterion. A prospective research agenda contains application of the idea of penalization of parameter heterogeneity in nonlinear problems, for likelihoods other than Gaussian, and for loss functions other than quadratic.

References

Baltagi, B., Bresson, G. and Pirotte, A. (2008) To pool or not to pool. Chapter 16 in *Panel Data Econometrics*, László Mátyás and Patrick Sevestre (eds), Springer, Heidelberg, pp. 517–546.

Baltagi, B.H. and Pinnoi, N. (1995) Public capital stock and state productivity growth: further evidence. *Empirical Economics*, 20(2), 351–359.

Caporin, M. & Paruolo, P. (2015). Proximity-structured multivariate volatility models. *Econometric Reviews*, 34(5), 559–593.

Engle, R.F., and Kroner, K.F. (1995) Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11, 122–150.

Golosnoy, V., Gribisch, B., and Liesenfeld, R. (2012) The conditional autoregressive Wishart model for multivariate stock market volatility. *Journal of Econometrics*, 167(1), 211–223.

Gruber, M.H.J. (2010) *Regression Estimators: A Comparative Study.* The John Hopkins University Press.

Hansen, B. (2016) Efficient shrinkage in parametric models. *Journal of Econometrics*, 190, 115–132.

Hoerl, A. E. (1959) Optimum solution of many variables equations. *Chemical Engineering Progress*, 55, 69–78.

Hoerl, A.E. and Kennard, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

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

Swindel, B.F. (1976) Good ridge estimators based on prior information. *Communications in Statistics – Theory and Methods*, 5(11), 1065–1075.

Theobald, C.M. (1974) Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36(1), 103–106.

van Wieringen, W.N. (2019) Lecture notes on ridge regression. `arXiv:1509.09169v4`.

Ziemer, R.F. and Wetzstein, M.E. (1983) A Stein-rule method for pooling data. *Economics Letters*, 11, 137–143.

Appendix

Proof of Proposition 1. First, let us compute

$$\begin{aligned}\hat{\beta}_{\text{ridge}} - \beta &= \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \sum_{t=1}^{n} x_t \left(x_t'\beta + e_t\right) - \beta \\ &= \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \sum_{t=1}^{n} x_t e_t - \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} n\lambda_\beta \Xi_P \beta,\end{aligned}$$

hence the conditional MSE $E\left[(\hat{\beta}_{\text{ridge}} - \beta)(\hat{\beta}_{\text{ridge}} - \beta)'|X\right]$ is

$$\begin{aligned}&\left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} E\left[\left(\sum_{t=1}^{n} x_t e_t\right)\left(\sum_{t=1}^{n} x_t e_t\right)' |X\right] \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \\ &\quad + \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \lambda_\beta^2 \Xi_P E\left[\beta\beta'\right] \Xi_P \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \\ &= \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \sigma^2 \left(\sum_{t=1}^{n} x_t x_t'\right) \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \\ &\quad + \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \lambda_\beta^2 \Xi_P \left(\beta_0^2 J_P + \omega_\beta I_P\right) \Xi_P \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1}.\end{aligned}$$

Now, the unconditional MSE $E\left[(\hat{\beta}_{\text{ridge}} - \beta)(\hat{\beta}_{\text{ridge}} - \beta)'\right]$ is

$$\begin{aligned}&E\left[\left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1} \left(\sigma^2 \sum_{t=1}^{n} x_t x_t' + \omega_\beta \lambda_\beta^2 \Xi_P\right) \left(\sum_{t=1}^{n} x_t x_t' + \lambda_\beta \Xi_P\right)^{-1}\right] \\ &= E\left[\left(I_P - \lambda_\beta \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \Xi_P + o_P\left(\frac{1}{n}\right)\right) \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \left(\sigma^2 \sum_{t=1}^{n} x_t x_t' + \omega_\beta \lambda_\beta^2 \Xi_P\right)\right. \\ &\qquad \left. \times \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \left(I_P - \lambda_\beta \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \Xi_P + o_P\left(\frac{1}{n}\right)\right)\right] \\ &= E\left[\left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \left(\sigma^2 \left(\sum_{t=1}^{n} x_t x_t'\right) + \omega_\beta \lambda_\beta^2 \Xi_P\right) \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1}\right] \\ &\qquad - 2\sigma^2 \lambda_\beta E\left[\left(\sum_{t=1}^{n} x_t x_t'\right)^{-1} \Xi_P \left(\sum_{t=1}^{n} x_t x_t'\right)^{-1}\right] + o\left(\frac{1}{n^2}\right).\end{aligned}$$

Collecting the terms and expanding further, we get

$$\begin{aligned}
&\frac{\sigma^2}{n}E\left[\left(E\left[x_t x_t'\right]+\frac{1}{n}\sum_{t=1}^{n}\left(x_t x_t'-E\left[x_t x_t'\right]\right)\right)^{-1}\right] \\
&\qquad+\left(\omega_\beta\lambda_\beta^2-2\sigma^2\lambda_\beta\right)E\left[\left(\sum_{t=1}^{n}x_t x_t'\right)^{-1}\Xi_P\left(\sum_{t=1}^{n}x_t x_t'\right)^{-1}\right]+o\left(\frac{1}{n^2}\right) \\
&=\frac{\sigma^2}{n}E\left[\left(I_P-E\left[x_t x_t'\right]^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(x_t x_t'-E\left[x_t x_t'\right]\right)\right.\right. \\
&\qquad\left.\left.+\left(E\left[x_t x_t'\right]^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(x_t x_t'-E\left[x_t x_t'\right]\right)\right)^2+o_P\left(\frac{1}{n}\right)\right)E\left[x_t x_t'\right]^{-1}\right] \\
&\qquad+\frac{\omega_\beta\lambda_\beta^2-2\sigma^2\lambda_\beta}{n^2}E\left[E\left[x_t x_t'\right]^{-1}\Xi_P E\left[x_t x_t'\right]^{-1}+o_P\left(1\right)\right]+o\left(\frac{1}{n^2}\right).
\end{aligned}$$

The asymptotic analog of this expression, omitting the remainder term, is equal to

$$\frac{\sigma^2}{n}I_P+\frac{\sigma^2}{n^2}\left(P+1\right)I_P+\frac{\omega_\beta\lambda_\beta^2-2\sigma^2\lambda_\beta}{n^2}\Xi_P.$$

The second term is due to the fact that

$$E\left(\frac{1}{n}\sum_{t=1}^{n}\left(x_t x_t'-I_P\right)\right)^2=\frac{1}{n}E\left(\left(x_t x_t'-I_P\right)\right)^2=\frac{1}{n}\left(P+1\right)I_P,$$

as the elements on the diagonal of $E\left(\left(x_t x_t'-I_P\right)\right)^2$ have expectation

$$E\left(\left(x_{jt}^2-1\right)^2+x_{jt}^2\sum_{i\neq j}x_{it}^2\right)=2+\left(P-1\right)=P+1,$$

and the off-diagonal elements have zero expectation.

Now, let (y_t^*, x_t^*) be drawn from the population of (y_t, x_t) independently from the given sample. The mean squared prediction error is

$$\begin{aligned}
MSE_{\text{ridge}} &= E\left[(y_t^*-x_t^{*\prime}\hat{\beta}_{\text{ridge}})^2\right] \\
&= E\left[(y_t^*-x_t^{*\prime}\beta)^2\right]+E\left[(\hat{\beta}_{\text{ridge}}-\beta)'x_t^*x_t^{*\prime}(\hat{\beta}_{\text{ridge}}-\beta)\right] \\
&= \sigma^2+\text{tr}\left(E\left[(\hat{\beta}_{\text{ridge}}-\beta)(\hat{\beta}_{\text{ridge}}-\beta)'\right]\right),
\end{aligned}$$

whose asymptotic analog is

$$\begin{aligned}
AMSE_{\text{ridge}} &= \sigma^2+\text{tr}\left(\frac{\sigma^2}{n}I_P+\frac{\sigma^2}{n^2}\left(P+1\right)I_P+\frac{\omega_\beta\lambda_\beta^2-2\sigma^2\lambda_\beta}{n^2}\Xi_P\right) \\
&= \sigma^2\left(1+\frac{P}{n}+\frac{P\left(P+1\right)}{n^2}\right)+\left(\omega_\beta\lambda_\beta^2-2\sigma^2\lambda_\beta\right)\frac{P-1}{n^2}.
\end{aligned}$$

□

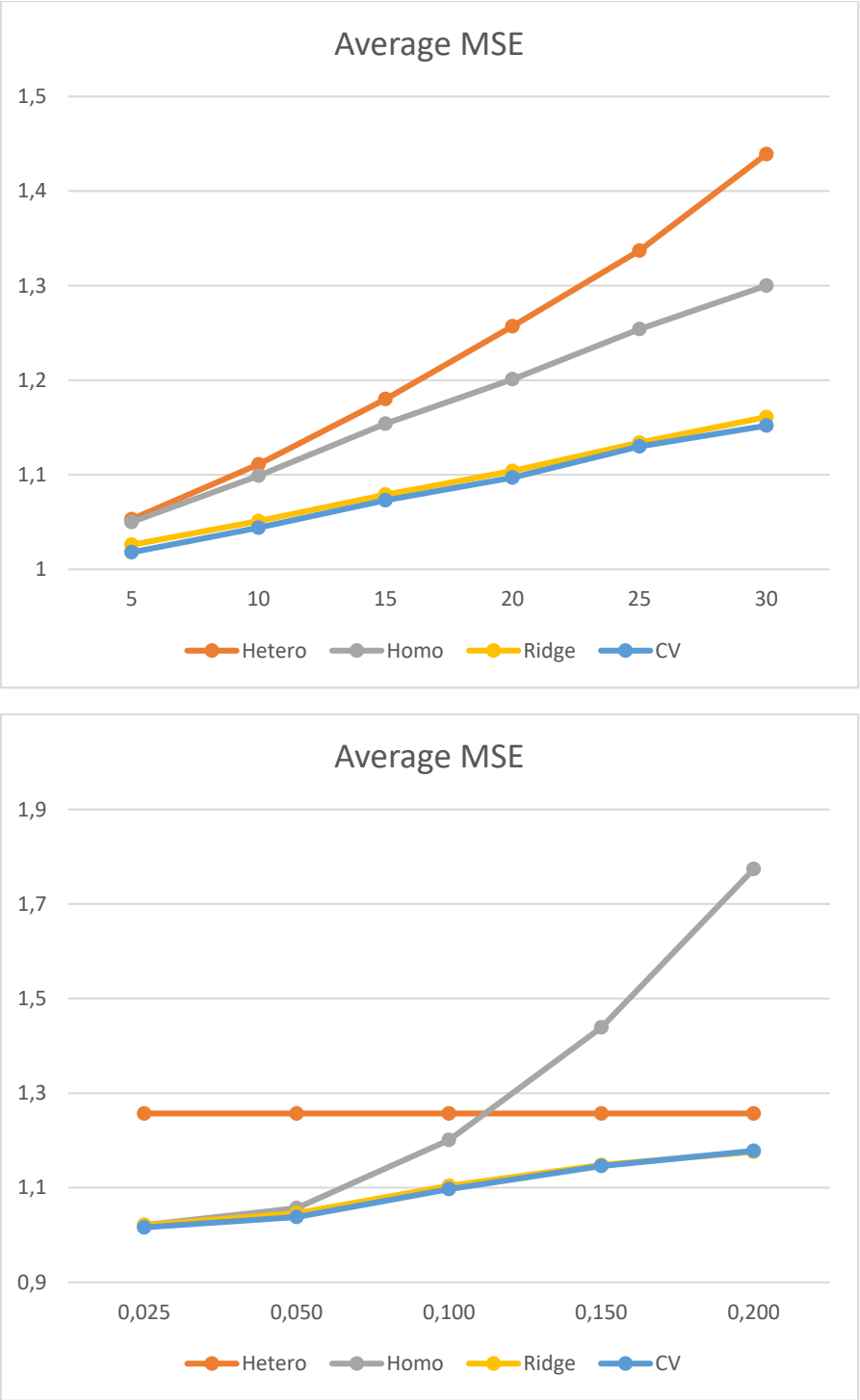

Figure 1. Graphs of dependence of average (across simulations) out of sample MSE on the degree of heterogeneity (top panel) and on parameter dimensionality (bottom panel), the one-dimensional prototype model.

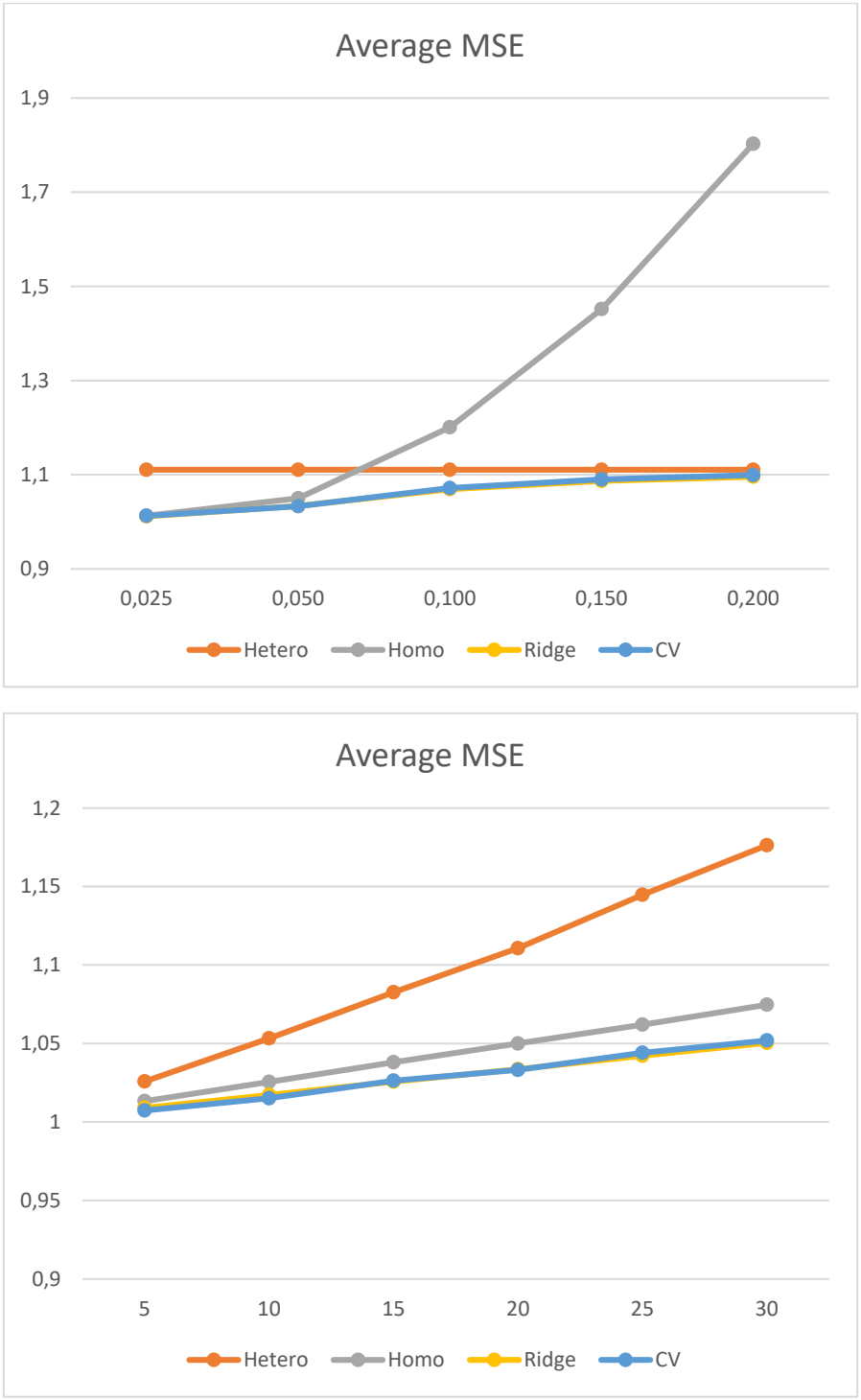

Figure 2. Graphs of dependence of average (across simulations) out of sample MSE on the degree of heterogeneity (top panel) and on parameter dimensionality (bottom panel), the two-dimensional prototype model with the fully homogeneous target.

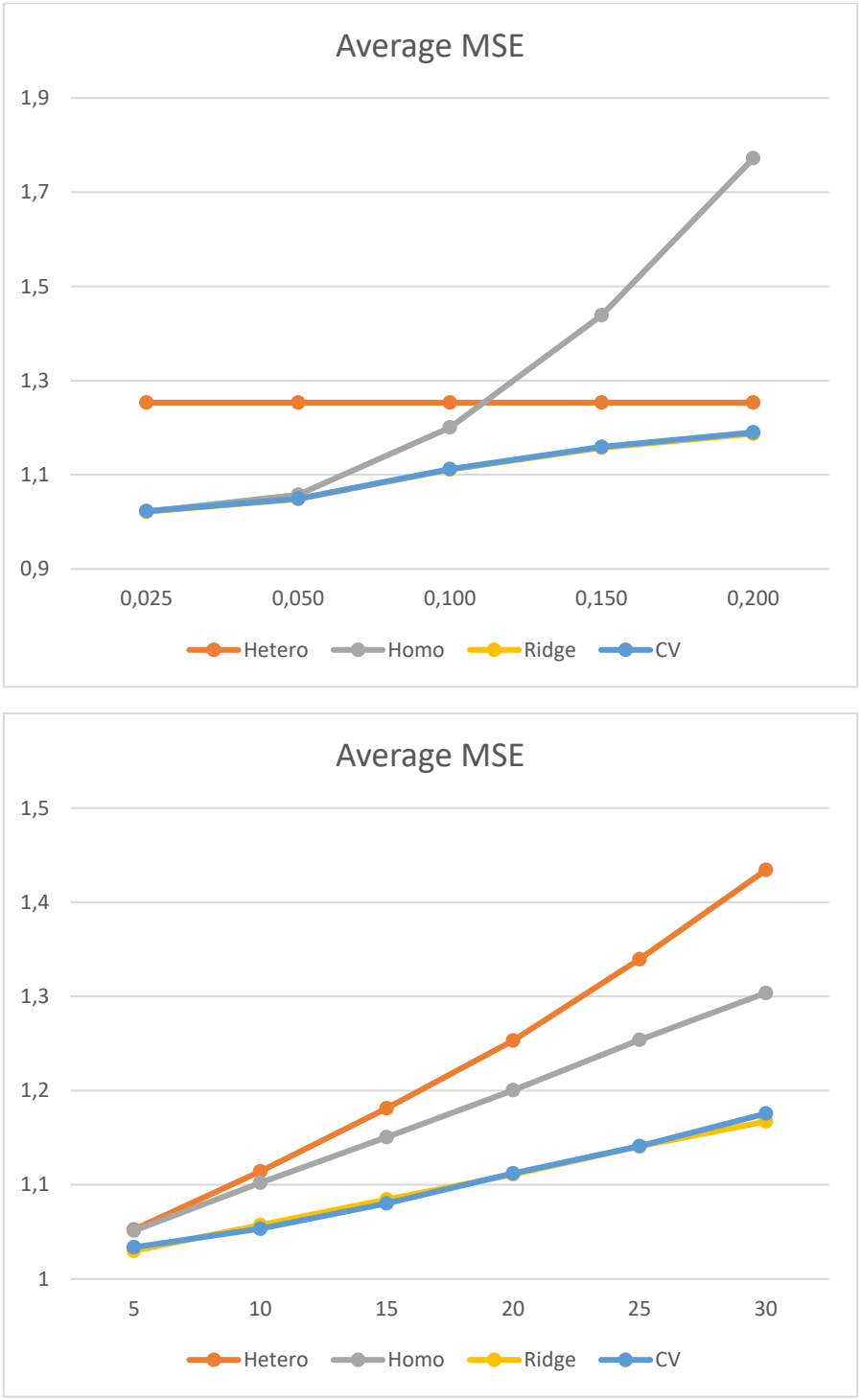

Figure 3. Graphs of dependence of average (across simulations) out of sample MSE on the degree of heterogeneity (top panel) and on parameter dimensionality (bottom panel), the two-dimensional prototype model with the row-wise homogeneous target.