Ridging out many covariates

Stanislav Anatolyev[*]
CERGE-EI

April 2025

Abstract

The paper considers a conditionally heteroskedastic linear regression setup with few regressors of interest and many nuisance covariates. We propose to subject the parameters corresponding to those nuisance covariates to a generalized ridge shrinkage. We show that under the assumption of dense random effects from the nuisance covariates, the *ridge-out* estimator of the parameters of interest is conditionally unbiased, and derive the optimal ridge intensity that delivers conditional efficiency. When tight structures on the variance of random effects are imposed, the asymptotic variance of the ridge-out estimator, under the dimension asymptotics, may be arbitrarily smaller than that of the least squares estimator. We also demonstrate how the optimal ridge-out estimator can be implemented under tight structures on the variance of random effects, and run simulation experiments where significant efficiency gains are possible to reach.

[*] Address: Stanislav Anatolyev, CERGE-EI, Politickych vězňů 7, 11121 Prague 1, Czech Republic; e-mail `stanislav.anatolyev@cerge-ei.cz`. This research was supported by the grant 24-12720S from the Czech Science Foundation.

1 Introduction

Applied statisticians often run 'long' regressions, where one is interested in one or maximum a handful of 'structural' coefficients, while throwing in a lot of other regressors whose coefficients are of at most secondary importance, or of no importance at all, in order to control all factors that may affect the outcome variable (Angrist and Hahn 2004, Cattaneo, Jansson, and Newey, 2018). When considered as a group, these 'nuisance' regressors (which we call *covariates*) may have a significant effect on the outcome as a group, but that effect comes from small effects of each of these covariates. These effects can be elegantly formalized as a hypothesis of *dense random effects* (Dicker and Erdogdu 2017, Dobriban and Wager, 2018, Liu and Dobriban, 2020) meaning that each covariate has a small (possibly independent) effect of the outcome variable.

As is well-known from the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933, Lovell, 1963), the least squares estimation of the coefficients of interest together with the nuisance coefficients is equivalent to first partialling out (in another terminology, filtering out) the nuisance regressors and running least squares of partialled-out outcomes on the partialled-out regressors. However, in the presence of many covariates with special structures imposed on their coefficients, other, more beneficial opportunities arise. We show that under the assumption of dense random effects, ridge-type shrinkage applied to the covariates, which results in a *ridge-out estimator* for the parameters of interest, leads to more efficient estimation than least squares. Notice that the interest to the ridge machinery has increased lately, especially in relation to high-dimensional environments, which in conjunction provide new opportunities – see, inter alia, Dicker (2016), van Wieringen and Peeters (2016), Park (2017), Dobriban and Wager (2018) and Anatolyev (2020); see also the survey by van Wieringen (2023).

For a conditionally heteroskedastic linear regression, we derive the form of optimal generalized ridge intensity, which turns out to have a particular non-trivial form, and a closed-form expression for the conditional variance. The optimal ridge-out estimator turns out to achieve the efficiency bound for linear unbiased estimation. Moreover, it is numerically equivalent to the GLS estimator of the coefficients of interest if the regression is viewed as a general mixed model (Robinson, 1991, Jiang, 1996), with the effects from covariates being the elements of the combined error. We allow the number of these effects, however, to asymptotically grow and, in particular, be proportional to the sample size.

For a couple of specific cases of random effects, we derive the limit of the variance ratio of the optimal ridge-out and least squares estimators under the dimension asymptotics. The variance ratio turns out to be able to take any values between 0 and 1 – that is, ridge-out is able to achieve large asymptotic efficiency gains compared to least squares when the numerosity of covariates is sizable, in a stark contrast to the conventional few-covariate asymptotic environment, in which the ridge machinery does not alter the asymptotic properties of estimators.

In the last part of the paper, we show how the optimal ridge-out estimation can be implemented, and verify its unbiasedness and relative efficiency in a small simulation study, where it turns out possible to attain significant efficiency gains. We also verify the ability of variance estimates to account for the actual variability of the ridge-out parameter estimates.

We emphasize that throughout the theory – except in the specific examples, – we do not place any distributional assumptions on any objects – not on the regression errors, not on the regressors of interest, not on the covariates, not on the random effects (beyond, of course, existence of necessary moments). The covariance structure of the regressors and/or covariates is not restricted either.

The paper is organized as follows. In section 2, we lay out the setup, describe the ridge-out estimator, and derive its properties, both generally in finite samples and asymptotically with certain structures imposed on the random effects. In section 3, we describe implementation under those certain structures, and study finite sample properties of the ridge-out estimator in simulations. Section 4 concludes. The Appendix contains technical derivations and proofs. A remark on notation: $\operatorname{tr}(A)$ denotes a trace of a square matrix A, and by $\lambda_j(A)$ we denote the j^{th} largest eigenvalue of a symmetric positive semidefinite matrix A.

2 Ridge-out estimator

2.1 Setup

For a random sample $\{(y_i, x_i, w_i)\}_{i=1}^n$, we consider the regression model

$$y_i = x_i'\beta_0 + w_i'\gamma_0 + e_i, \tag{1}$$

where β_0 is $k \times 1$ vector of structural coefficients, γ_0 is $m \times 1$ vector of other coefficients, and

$$E[e_i | x_i, w_i] = 0.$$

Most of our results hold in the general case of conditional heteroskedasticity, where the individual variances $\sigma_i^2 = E[e_i^2 | x_i, w_i]$ vary with observations. In the examples, we impose conditional homoskedasticity, $\sigma_i^2 = \sigma^2$ for all $i = 1, ..., n$, in order to obtain clear-cut insights. Note that we do not impose any distributional assumptions on the error term, apart from existence of conditional moments.

As can be seen from (1), the set of right side variables is partitioned into two groups: x_i are the (few) *regressors*, while w_i are the (possibly many) *covariates*. The partition into the two groups is such that the vector β_0 collects (i) the coefficients of interest (often, a single coefficient) corresponding to the factors of interest, and (ii) those coefficients that correspond to other influential factors. In contrast, the vector γ_0 collects the marginal effects for nuisance covariates, which are of no interest, which are numerous, and which have a significant influence on the outcome variable only as a group. Our framework allows unrestricted covariance structures among the regressors, among the covariates, and between the two groups.

In a matrix notation,

$$Y = X\beta_0 + W\gamma_0 + e, \tag{2}$$

where $Y = (y_1, ..., y_n)'$ is $n\times 1$, $X = (x_1, ..., x_n)'$ is $n\times k$, $W = (w_1, ..., w_n)'$ is $n\times m$, and $e = (e_1, ..., e_n)'$ is $n \times 1$. We assume that the rank of matrix (X, W) equals its column dimension $k + m$. Denote for future use the projection and annihilation matrices $P_W = W(W'W)^{-1} W'$, $M_W = I_n - P_W$, $P_X = X(X'X)^{-1} X'$ and $M_X = I_n - P_X$. Due to random sampling, the error vector satisfies $E[e|X, W] = 0$ and $E[ee'|X, W] = \text{diag}\left\{\sigma_i^2\right\}_{i=1}^n \equiv \Sigma$.

We impose the following random effects design for the covariates. It means that each covariate has a small random effect on the outcome variable, see Dobriban and Wager (2018). Initially we make it more general than that in Dobriban and Wager (2018), and make it as restrictive when confronted with practical implementation.

Assumption 1 *Assume the dense random effects covariate design:* γ_0 *is a random vector with* $E(\gamma_0|X, W) = 0$ *and var* $(\gamma_0|X, W) = \Gamma$, *conditionally independent of* e.

Note that we do not impose any distributional assumptions on γ_0 either.

2.2 Ridge-out estimation

We consider a generalized ridge regression estimator where only the γ coefficients are penalized.

$$\begin{pmatrix} \hat{\beta}_\Xi(\Xi) \\ \hat{\gamma}_\Xi(\Xi) \end{pmatrix} = \begin{bmatrix} X'X & X'W \\ W'X & \Xi \end{bmatrix}^{-1} \begin{bmatrix} X' \\ W' \end{bmatrix} Y, \tag{3}$$

where Ξ is a symmetric positive definite $m \times m$ ‘ridge’ matrix that is a function of only (X, W). The least squares estimator of β_0 corresponds to using $\Xi = W'W$; let us denote it by $\hat{\beta}_{LS} = \hat{\beta}_\Xi(W'W)$. We call the estimator $\hat{\beta}_\Xi(\Xi)$ of β_0 a *ridge-out* estimator, a hybrid of the partialled-out least squares estimator $\hat{\beta}_{LS}$ and the regular generalized ridge estimator (Hoerl and Kennard, 1970) when only the γ-coefficients are penalized.[1] The ridge matrix Ξ generalizes the form of the classical ridge of $\Xi = W'W - \lambda_W I_m$, for a scalar penalty parameter $\lambda_W > 0$.

2.3 Properties of ridge-out

It turns out that for any legitimate choice of ridge intensity Ξ, the ridge-out estimator is unbiased under the assumption of random effects.

Proposition 1 *Under Assumption 1, $\hat{\beta}_\Xi(\Xi)$ is conditionally unbiased.*

The unbiasedness of the ridge-out estimator is a remarkable property. Recall that the standard ridge machinery introduces bias for all the coefficient estimates, this bias being to be traded for a smaller variance. The ridge-out method instead keeps the estimates of the structural parameters unbiased (like the least squares does), and the variance reduction is achieved at the expense of biasedness of the nuisance coefficient estimates only, those that are of no interest.

The following proposition provides an optimal choice of ridge intensity in the sense of minimal conditional variance of the ridge-out estimates.

Proposition 2 *Under Assumption 1, the conditional variance of $\hat{\beta}_\Xi(\Xi)$ is equal to*

$$var\left(\hat{\beta}_\Xi(\Xi)\,|X,W\right) = \left(X'\Psi_\Xi X\right)^{-1} X'\Psi_\Xi\left(W\Gamma W' + \Sigma\right)\Psi_\Xi X\left(X'\Psi_\Xi X\right)^{-1}, \tag{4}$$

where $\Psi_\Xi = I_n - W\Xi^{-1}W'$, and is minimized by choosing

$$\Xi^* = W'\left(I_n - \left(P_W - W\Gamma W' - \Sigma\right)^{-1}\right)W.$$

The minimal conditional variance is

$$var\left(\hat{\beta}_\Xi(\Xi^*)\,|X,W\right) = \left(X'\left(W\Gamma W' + \Sigma\right)^{-1}X\right)^{-1}. \tag{5}$$

Note that when no ridging is performed, $\Xi = W'W$ and $\Psi_\Xi = M_W$, which simply leads to the conditional variance of least squares, which is suboptimal,

$$var\left(\hat{\beta}_{LS}|X,W\right) = \left(X'M_W X\right)^{-1} X'M_W\Sigma M_W X\left(X'M_W X\right)^{-1}, \tag{6}$$

[1] The whole estimator $\left(\hat{\beta}_\Xi(\Xi)', \hat{\gamma}_\Xi(\Xi)'\right)'$ can also be viewed as a generalized ridge estimator with a constraint that no ridging of the β-subvector is done.

and so (6) is generally larger than (5). Evidently, one may do better in terms of conditional variance than partialling out least squares implied by running least squares, keeping unbiasedness intact at the same time. The optimal ridging shrinks the least squares-implied block $W'W$ by the matrix $W'\left(P_W - W\Gamma W' - \Sigma\right)^{-1} W$, which evidently has a more complex structure than shrinkage towards the identity matrix implied by the classical ridge. The optimal amount of ridging depends, in addition to realization of covariates, on the conditional error variance matrix Σ and the explanatory power of covariates $W\Gamma W'$. Interestingly, when $\gamma_0 = 0$ and so $\Gamma = 0$, the optimal ridge-out does apply shrinkage to $W'W$ and is thus still superior to least squares in terms of conditional efficiency.

2.4 General mixed model perspective

The model (2) under Assumption 1 can be viewed as a general mixed model (e.g., Robinson, 1991, Jiang, 1996)

$$Y = X\beta_0 + u,$$

where $u = W\gamma_0 + e$ with the properties $E\left[u|X,W\right] = 0$ and $var\big(u|X,W\big) = W\Gamma W' + \Sigma$. The efficient (best linear unbiased) estimator of β for this model is

$$\hat{\beta}_{GLS} = \left(X'\left(W'\Gamma W + \Sigma\right)^{-1} X\right)^{-1} X'\left(W\Gamma W' + \Sigma\right)^{-1} Y,$$

with conditional variance

$$var\big(\hat{\beta}_{GLS}|X,W\big) = \left(X'\left(W\Gamma W' + \Sigma\right)^{-1} X\right)^{-1},$$

which coincides with $var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)$. The next result then follows.

Proposition 3 *Under Assumption 1, the optimal ridge-out estimator is efficient in the class of linear unbiased estimators.*

Moreover, it turns out that the ridge-out and GLS estimators are numerically equivalent.

Proposition 4 *The optimal ridge-out and GLS estimators are equal: $\hat{\beta}_{\Xi}\left(\Xi^*\right) = \hat{\beta}_{GLS}$.*

2.5 Structured random effects

Most interesting cases of the random design are the following two special structures of Γ.

- **Design A.** The special case of random effects design is $\Gamma = \frac{\alpha^2}{m} I_m$.
- **Design B.** The special case of random effects design is $\Gamma = \alpha^2 \left(W'W\right)^{-1}$.

In Design A, the effects of different covariates are uncorrelated but depend on their values; in Design B, orthonormalization is invoked to equalize the impact across the covariates. Here, α^2 indexes the strength of the signal, the signal being uniformly distributed across the m covariates. In the case of Design A, the optimal ridge-out equals

$$\Xi^* = W' \left(I_n - \left(P_W - \frac{\alpha^2}{m} WW' - \Sigma \right)^{-1} \right) W,$$

so that

$$var\big(\hat{\beta}_\Xi \left(\Xi^*\right) | X, W\big) = \left(X' \left(\frac{\alpha^2}{m} WW' + \Sigma \right)^{-1} X \right)^{-1}.$$

In the case of Design B, the optimal ridge-out equals

$$\Xi^* = W'\big(I_n - \big((1-\alpha^2)P_W - \Sigma\big)^{-1}\big)W,$$

so that

$$var\big(\hat{\beta}_\Xi \left(\Xi^*\right) | X, W\big) = \left(X' \left(\alpha^2 P_W + \Sigma \right)^{-1} X \right)^{-1}.$$

Again, being optimal in the class of generalized ridge estimators, the optimal ridge-out estimator is conditionally more efficient than least squares.

2.6 Examples and asymptotics

In order to quantify relative efficiency, we consider a simple example, where it is possible to find the limiting variances of both least squares and ridge-out estimators in a closed form, and compare them in terms of asymptotic efficiency. To this end, first, we impose conditional homoskedasticity, $\Sigma = \sigma^2 I_n$. Second, we employ the dimension asymptotics, which is characteristic of the random matrix theory (e.g., see Bai and Silverstein, 2010) and regression theory with many regressors (e.g., see Anatolyev, 2012):

Assumption 2 *$m \to \infty$ along with $n \to \infty$, such that $m/n = \mu + o(n^{-1/2})$, where $0 \le \mu < 1$, while k stays fixed.*

The conventional asymptotics obtains for $\mu = 0$.

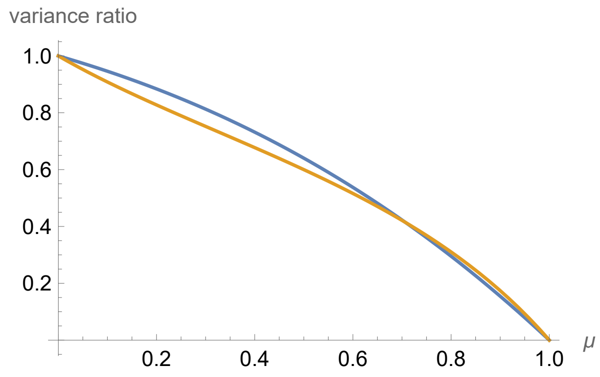

Figure 1: Figure 1. Asymptotic variance ratio as a function of μ.

Set $k = 1$. Let the only column of X contain i.i.d. zero mean unit variance random variables, and each of m columns of W contain i.i.d. zero mean unit variance random variables that are also independent of X.[2] Denote

$$\omega = \frac{\alpha^2}{\mu\sigma^2}.$$

Proposition 5 *Under the dimension asymptotics of Assumption 2, we have, for the least squares and optimal ridge-out, respectively,*

$$\frac{var\left(\hat{\beta}_{\Xi}\left(\Xi^{*}\right)|X,W\right)}{var\left(\hat{\beta}_{LS}|X,W\right)} \xrightarrow{P} (1-\mu) \times \begin{cases} \dfrac{2}{\sqrt{\left(\omega^{-1}+\mu-1\right)^{2}+4\omega^{-1}}-\left(\omega^{-1}+\mu-1\right)} & \textit{under Design A,} \\ \left(1-\dfrac{\mu^{2}}{\omega^{-1}+\mu}\right)^{-1} & \textit{under Design B.} \end{cases}$$

The variance ratio depends on α^2 via ω^{-1}. Under Design A, if the signal strength vanishes, $\alpha^2 \to 0$, then $\lim n \cdot var\left(\hat{\beta}_{\Xi}\left(\Xi^{*}\right)|X,W\right) = \sigma^2 < \lim n \cdot var\left(\hat{\beta}_{LS}|X,W\right) = (1-\mu)^{-1}\sigma^2$, and the inequality may be very loose when the dimension ratio μ is large. Set $\alpha^2 = \mu$ and $\sigma^2 = 1$. Then $\omega = 1$ and

$$\frac{var\left(\hat{\beta}_{\Xi}\left(\Xi^{*}\right)|X,W\right)}{var\left(\hat{\beta}_{LS}|X,W\right)} \xrightarrow{P} \frac{2\left(1-\mu\right)}{\sqrt{\mu^{2}+4}-\mu},$$

under Design A, and

$$\frac{var\left(\hat{\beta}_{\Xi}\left(\Xi^{*}\right)|X,W\right)}{var\left(\hat{\beta}_{LS}|X,W\right)} \xrightarrow{P} \frac{1-\mu^{2}}{1+\mu-\mu^{2}}$$

under Design B. The graphs of the limiting variance ratios are depicted on Figure 1, in blue for Design A, and in orange for Design B.

[2] Independence of X and W is needed to obtain clean closed form formulas for the asymptotic variance ratio.

The limiting variance ratio takes all values on $[0,1]$. Under either design, when covariates are few so that $\mu = 0$, ridging them out is asymptotically negligible,[3] as the regular ridge results only in finite sample changes. Under either design, if covariates are many and $0 < \mu < 1$, returns from ridging-out are monotonically increasing in μ, and for big μ the ridging-out can be arbitrarily more asymptotically efficient than least squares.

3 Implementation

3.1 Feasible ridge-out

We try to outline implementation of the ridge-out estimator under Design A and Design B. A strong assumption of this kind is needed to reduce degrees of freedom from too many in a general Γ. We also need $\mu > 0$. Implementation in practice is worthwhile only if one is confident of applicability of the random effects assumption, as its violation will lead to an estimation bias.

To implement the ridge-out estimator, one needs reliable estimates of σ^2 under homoskedasticity or Σ under heteroskedasticity, on the one hand, and of α^2, on the other. Let the full projection matrix be $P = (X,W)\left((X,W)'(X,W)\right)^{-1}(X,W)'$ with the diagonal elements $\{P_{ii}\}_{i=1}^n$ and the corresponding annihilation matrix be $M = I_n - P$ with the diagonal elements $\{M_{ii}\}_{i=1}^n$. Further, let $\hat{e} = MY$ be the vector of least squares residuals with elements $\{\hat{e}_i\}_{i=1}^n$. An estimate of σ^2 can be easily constructed as a sample variance of residuals:

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-m-k}.$$

Note that this estimator is exactly conditionally unbiased. Under conditional heteroskedasticity, we take advantage of estimates of individual variances from Kline, Saggio and Sølvsten (2020), which are robust to covariate numerosity:

$$\hat{\sigma}_i^2 = \frac{y_i\hat{e}_i}{M_{ii}}.$$

These estimates originate from leave-one-out LS estimation and equality between LS residuals $\hat{e}_i$ corrected for leverage and leave-one-out LS residuals. Importantly, these estimates are exactly condi-

[3] Note that with no restrictions on α^2, as $\mu \to 0$ and α^2 is fixed, $\omega^{-1} \to 0$ and the limit of $n\,var\big(\hat{\beta}_\Xi\left(\Xi^*\right)|X,W\big)$ is σ^2.

tionally unbiased, and this property holds for any number of covariates once the rank condition holds.[4] We construct the error variance matrix estimator as

$$\hat{\Sigma} = \begin{cases} \hat{\sigma}^2 I_n & \text{under conditional homoskedasticity,} \\ \text{diag}\left\{\hat{\sigma}_i^2\right\}_{i=1}^n & \text{under conditional heteroskedasticity.} \end{cases}$$

The quantity α^2 can be identified from the variation across nuisance coefficients $\gamma_0'\gamma_0$, and estimated from the sample variability of the coefficient estimates. Observe that for the least squares estimates,

$$\hat{\gamma} = \gamma_0 + \left(W'M_X W\right)^{-1} W'M_X e. \tag{7}$$

For Design A, we form quadratic forms of both sides of (7) with respect to the identity matrix, and take expectations:

$$E\left[\hat{\gamma}'\hat{\gamma}\right] = E\left[\gamma_0'\gamma_0\right] + E\left[\text{tr}\big((W'M_X W)^{-1}W'M_X \Sigma M_X W(W'M_X W)^{-1}\big)\right].$$

Under Design A, $E\left[\gamma_0'\gamma_0\right] = \text{tr}\left(E\left[\gamma_0\gamma_0'\right]\right) = \text{tr}\left(\Gamma\right) = \alpha^2$. Plugging in the sample analogs – $\hat{\gamma}'\hat{\gamma}$ for the left side and $\hat{\Sigma}$ for Σ in the right side, – we obtain an equation defining the estimate of α^2:

$$\hat{\alpha}^2 = \hat{\gamma}'\hat{\gamma} - \text{tr}\big((W'M_X W)^{-1}W'M_X \hat{\Sigma} M_X W(W'M_X W)^{-1}\big).$$

For Design B, we form quadratic forms of both sides of (7) with respect to the matrix $W'W$, and take expectations:

$$E\left[\hat{\gamma}'\left(W'W\right)\hat{\gamma}\right] = E\left[\gamma_0'\left(W'W\right)\gamma_0\right] + E\left[\text{tr}\big((W'M_X W)^{-1}W'M_X \Sigma M_X W(W'M_X W)^{-1}\left(W'W\right)\big)\right].$$

Under Design B, $E\left[\gamma_0'\left(W'W\right)\gamma_0\right] = \text{tr}\left(E\left[W'WE\left[\gamma_0\gamma_0'|W\right]\right]\right) = \text{tr}\left(E\left[W'W\Gamma\right]\right) = \alpha^2\text{tr}\left(I_m\right) = m\alpha^2$. Plugging in the sample analogs – $\hat{\gamma}'\left(W'W\right)\hat{\gamma}$ for the left side and $\hat{\Sigma}$ for Σ in the right side, – we obtain an equation defining the estimate of α^2:

$$\hat{\alpha}^2 = \frac{\hat{\gamma}'\left(W'W\right)\hat{\gamma} - \text{tr}\big((W'M_X W)^{-1}W'M_X \hat{\Sigma} M_X W(W'M_X W)^{-1}(W'W)\big)}{m}.$$

[4] The unbiasedness property follows because, due to conditional mean zero errors and random sampling,

$$\begin{aligned} E\left[\hat{\sigma}_i^2|X,W\right] &= M_{ii}^{-1}E\big[\left(x_i'\beta_0 + w_i'\gamma_0 + e_i\right)\sum_{j=1}^n M_{ij}e_j|X,W\big] \\ &= M_{ii}^{-1}\left(x_i'\beta_0 + w_i'\gamma_0\right)E\big[\sum_{j=1}^n M_{ij}e_j|X,W\big] + M_{ii}^{-1}E\big[e_i\sum_{j=1}^n M_{ij}e_j|X,W\big] \\ &= M_{ii}^{-1}\sum_{j=1}^n M_{ij}E\left[e_ie_j|X,W\right] = \sigma_i^2. \end{aligned}$$

Note that under both designs, the second term in $\hat{\alpha}^2$ is an exactly unbiased estimate of its expectation because we use an estimate $\hat{\Sigma}$ that is exactly conditionally unbiased. Then, we set the feasible ridge-out correspond to using $\hat{\sigma}^2 I_n$ in construction of $\hat{\alpha}^2$ in the homoskedastic case and $\hat{\Sigma}$ in construction of $\hat{\alpha}^2$ in the heteroskedastic case, and using $\hat{\sigma}^2 I_n$ in construction of Ξ^* and $var\left(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X, W\right)$.

3.2 Simulation evidence

We set $\beta = 1$, $\sigma^2 = 1$ and vary α^2 in the set $\{0, 0.5, 2.0, 5.0\}$. The main regressors are drawn independently from $\mathcal{N}(0,1)$. We generate $\frac{1}{5}$ of the nuisance regressors from $\mathcal{B}\left(\frac{1}{2}\right)$, another $\frac{1}{5}$ from $\mathcal{U}[0,1]$, yet another $\frac{1}{5}$ from $\mathcal{N}(0,1)$, yet another $\frac{1}{5}$ from $\chi^2(1)$, and the last $\frac{1}{5}$ from $\mathcal{LN}(0,1)$, independently from each other. We intentionally do not normalize the nuisance regressors so that the mean and variances may vary across them. The errors are generated by $e_i = \sigma_i \eta_i$, where η_i are drawn independently from $\mathcal{N}(0,1)$; in the homoskedastic case, $\sigma_i = 1$, and in the heteroskedastic case, $\sigma_i = \sqrt{nP_{ii}/(k+m)}$. Note that in both cases, the error variance is unity.[5] We vary n in the set $\{100, 200, 400\}$ and m/n in the set $\left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$. Because $\hat{\alpha}^2$ is not guaranteed to be positive, in unfavorable circumstances with small sample sizes when $\hat{\alpha}^2$ turns out negative, we replace the ridge-out estimator by the least squares one. All figures are based on 5,000 simulation runs.

The results are reported in Table 1 for Design A and in Table 2 for Design B, where we report the root mean squared error (RMSE) of least squares and feasible ridge-out estimators. We do not report estimation biases, as they are negligibly small, typically in the range $0.0005 \div 0.002$, so all RMSE figures effectively coincide with corresponding standard deviations. The tiny biases of the feasible ridge-out estimator are consistent with the unbiasedness property of its infeasible version (cf. Proposition 1). One can see from the tables that for all parameter and design combinations, there are visible efficiency gains from ridging out, monotonically and sharply increasing with m/n. In percentage terms, they are also higher for higher n; these percentages reach as high as almost 50% at times. The efficiency gains are very similar in the homoskedastic and heteroskedastic designs, the differences being tiny despite much more noisy estimation of Σ in the latter case. Interestingly, the efficiency of least squares is flat with respect to values of α^2, while the variance of the ridge-out estimator tends to be U-shaped, less pronounced for Design A and more pronounced for Design B.

We also verify the ability of variance estimates to account for the actual variability of the ridge-out parameter estimates. To this end, we construct a t-statistic based on the ridge-out estimates

[5] In the latter case, from the identity $\sum_{i=1}^{n} P_{ii} = \mathrm{rk}(P)$ and symmetry, $E\left[\sigma_i^2\right] = E\left[\mathrm{rk}(P)\right]/(k+m) = 1$.

and conditional variance estimates computed from (5). We approximate the critical values by the quantiles of the normal distribution.[6] In Table 3, we report rejection rates exhibited by the ridge-out t-statistic, corresponding to the 5% nominal size, for one moderate and one extreme values of α^2. The size distortions turns out to be small, despite pretty small sample sizes, not exceeding 1% in most cases except extremal ones; they are comparable in the homoskedastic and heteroskedastic cases and pretty flat with respect to small and moderate values of α^2. Towards combinations of high α^2 and high μ, the distortions kick in, and the actual size may reach twice the nominal size, though only for Design B.

4 Conclusion

When the regression contain many covariates of no interest, it is convenient and useful to apply the ridge machinery to these covariates. Under the hypothesis of dense random effects imposed on these covariates, the ridge-out is able to achieve a higher efficiency of estimation than that of least squares, while the optimal ridge-out estimator turns out to be best linear unbiased. In stylized examples, the efficiency gains are approximately proportional to the dimensionality of covariates. The optimal ridge-out methodology may be implemented in practice exploiting the variation across coefficient estimators of the nuisance covariates. Simulation outcomes show that in practice, the efficiency gains relative to least squares may well be large, while the size distortions of hypothesis tests based on the t-statistic tend to be small.

[6] The derivation of the asymptotic distribution of the ridge-out estimator is beyond the scope of this paper. This asymptotic distribution under suitable conditions is expected to be normal, as suggested by the literature on regression analysis with many regressors under dimension asymptotics (e.g., Anatolyev, 2012; Cattaneo, Jansson, and Newey, 2019). Asymptotic normality is pretty evident from the representation of $\hat{\beta}_{\Xi}(\Xi) - \beta_0$ as a sum of weighted averages of γ_0 and e.

Table 1. Root mean squared error of least squares and feasible optimal ridge-out estimators, DGP of Design A.

		$\alpha^2 = 0$	$\alpha^2 = 0.5$	$\alpha^2 = 2.0$	$\alpha^2 = 5.0$	$\alpha^2 = 0$	$\alpha^2 = 0.5$	$\alpha^2 = 2.0$	$\alpha^2 = 5.0$
		homoskedasticity				heteroskedasticity			
						$n = 100$			
least squares	$m = 25$	0.118	0.117	0.118	0.119	0.118	0.119	0.118	0.116
ridge-out		0.111	0.111	0.113	0.115	0.112	0.114	0.115	0.114
least squares	$m = 50$	0.147	0.144	0.147	0.144	0.144	0.144	0.144	0.146
ridge-out		0.130	0.125	0.127	0.128	0.127	0.127	0.128	0.129
least squares	$m = 75$	0.208	0.210	0.208	0.207	0.205	0.204	0.207	0.210
ridge-out		0.166	0.163	0.160	0.156	0.165	0.159	0.161	0.160
						$n = 200$			
least squares	$m = 50$	0.082	0.082	0.082	0.082	0.083	0.083	0.083	0.083
ridge-out		0.078	0.078	0.079	0.080	0.079	0.079	0.080	0.081
least squares	$m = 100$	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
ridge-out		0.088	0.086	0.086	0.090	0.088	0.087	0.088	0.089
least squares	$m = 150$	0.144	0.144	0.144	0.143	0.143	0.143	0.143	0.143
ridge-out		0.116	0.114	0.110	0.106	0.116	0.114	0.109	0.109
						$n = 400$			
least squares	$m = 100$	0.057	0.059	0.058	0.058	0.058	0.058	0.058	0.058
ridge-out		0.054	0.056	0.056	0.056	0.055	0.054	0.056	0.057
least squares	$m = 200$	0.072	0.072	0.071	0.072	0.071	0.071	0.071	0.070
ridge-out		0.062	0.061	0.061	0.064	0.062	0.061	0.061	0.063
least squares	$m = 300$	0.101	0.102	0.100	0.100	0.102	0.102	0.101	0.100
ridge-out		0.082	0.079	0.074	0.074	0.081	0.079	0.074	0.072

Table 2. Root mean squared error of least squares and feasible optimal ridge-out estimators, DGP of Design B.

		$\alpha^2 = 0$	$\alpha^2 = 0.5$	$\alpha^2 = 2.0$	$\alpha^2 = 5.0$	$\alpha^2 = 0$	$\alpha^2 = 0.5$	$\alpha^2 = 2.0$	$\alpha^2 = 5.0$
		homoskedasticity				heteroskedasticity			
$n = 100$									
least squares	$m = 25$	0.118	0.117	0.118	0.119	0.118	0.119	0.119	0.120
ridge-out		0.109	0.106	0.112	0.115	0.111	0.110	0.114	0.117
least squares	$m = 50$	0.147	0.144	0.144	0.144	0.143	0.146	0.144	0.145
ridge-out		0.123	0.114	0.125	0.132	0.120	0.114	0.127	0.135
least squares	$m = 75$	0.208	0.210	0.211	0.207	0.205	0.210	0.210	0.206
ridge-out		0.148	0.126	0.146	0.166	0.147	0.126	0.151	0.167
$n = 200$									
least squares	$m = 50$	0.082	0.082	0.082	0.082	0.083	0.083	0.083	0.083
ridge-out		0.076	0.075	0.078	0.080	0.077	0.076	0.079	0.081
least squares	$m = 100$	0.100	0.100	0.102	0.100	0.101	0.101	0.101	0.100
ridge-out		0.085	0.077	0.088	0.092	0.086	0.079	0.089	0.093
least squares	$m = 150$	0.144	0.144	0.143	0.143	0.142	0.144	0.145	0.143
ridge-out		0.107	0.084	0.099	0.115	0.108	0.084	0.106	0.117
$n = 400$									
least squares	$m = 100$	0.057	0.059	0.058	0.058	0.058	0.058	0.058	0.057
ridge-out		0.053	0.054	0.055	0.056	0.054	0.053	0.055	0.056
least squares	$m = 200$	0.072	0.072	0.071	0.072	0.073	0.071	0.071	0.072
ridge-out		0.061	0.055	0.062	0.067	0.062	0.055	0.063	0.067
least squares	$m = 300$	0.101	0.102	0.100	0.100	0.102	0.101	0.100	0.100
ridge-out		0.076	0.058	0.071	0.082	0.077	0.077	0.072	0.081

Table 3. Actual test sizes corresponding to 5% nominal size for feasible optimal ridge-out estimator.

	$\alpha^2 = 0.5$	$\alpha^2 = 5.0$	$\alpha^2 = 0.5$	$\alpha^2 = 5.0$	$\alpha^2 = 0.5$	$\alpha^2 = 5.0$	$\alpha^2 = 0.5$	$\alpha^2 = 5.0$
	Design A				Design B			
	homoskedasticity		heteroskedasticity		homoskedasticity		heteroskedasticity	
$n = 100$								
$m = 25$	6.1%	5.1%	6.7%	6.7%	5.2%	5.2%	6.1%	6.6%
$m = 50$	5.0%	5.3%	5.8%	6.1%	5.1%	6.2%	5.6%	6.9%
$m = 75$	4.3%	6.5%	5.0%	6.6%	5.8%	10.6%	6.1%	10.9%
$n = 200$								
$m = 50$	5.3%	5.3%	6.1%	6.0%	4.5%	5.3%	4.6%	5.8%
$m = 100$	4.7%	5.3%	5.1%	6.1%	4.6%	5.7%	5.1%	6.5%
$m = 150$	4.3%	6.3%	4.2%	6.1%	5.2%	9.9%	5.6%	10.2%
$n = 400$								
$m = 100$	4.8%	5.5%	6.0%	5.6%	4.0%	5.2%	4.9%	5.2%
$m = 200$	5.4%	4.8%	5.6%	5.7%	4.5%	5.5%	4.9%	6.3%
$m = 300$	4.4%	6.5%	4.3%	6.4%	5.3%	10.6%	5.1%	10.0%

A Appendix

Denote $\Upsilon_{\Xi} = \left(X'X - X'W\Xi^{-1}W'X\right)^{-1}$.

Lemma 1. The following equality holds:

$$\Upsilon_{\Xi}X'W\Xi^{-1} = \left(X'X\right)^{-1} X'W \left(\Xi - W'P_X W\right)^{-1}. \tag{8}$$

Proof of Lemma 1. Premultiply the identity

$$\Xi - W'P_X W = \Xi - W'P_X W.$$

by $X'W\Xi^{-1}$ to get

$$X'W\Xi^{-1}\left(\Xi - W'P_X W\right) = X'W - X'W\Xi^{-1}W'P_X W.$$

Represent $X'W$ on the right side as $X'P_X W$, then

$$X'W\Xi^{-1}\left(\Xi - W'P_X W\right) = \left(X'X - X'W\Xi^{-1}W'X\right)\left(X'X\right)^{-1} X'W.$$

Premultiplying by Υ_{Ξ} and postmultiplying by $(\Xi - W'P_X W)^{-1}$, we get

$$\Upsilon_{\Xi}X'W\Xi^{-1} = \left(X'X\right)^{-1} X'W \left(\Xi - W'P_X W\right)^{-1}$$

as stated. □

Proof of Proposition 1. Plugging in the model, we get

$$\begin{pmatrix} \hat{\beta}_{\Xi}(\Xi) \\ \hat{\gamma}_{\Xi}(\Xi) \end{pmatrix} = \begin{bmatrix} X'X & X'W \\ W'X & \Xi \end{bmatrix}^{-1} \left(\begin{bmatrix} X'X\beta_0 + X'W\gamma_0 \\ W'X\beta_0 + W'W\gamma_0 \end{bmatrix} + \begin{bmatrix} X'e \\ W'e \end{bmatrix} \right).$$

By the partitioned matrix inverse formula,

$$\begin{aligned} \hat{\beta}_{\Xi}(\Xi) &= \begin{bmatrix} \Upsilon_{\Xi} & -\Upsilon_{\Xi}X'W\Xi^{-1} \end{bmatrix} \begin{bmatrix} X' \\ W' \end{bmatrix} Y \\ &= \Upsilon_{\Xi}X'\left(I_n - W\Xi^{-1}W'\right)Y \\ &= \beta_0 + \Upsilon_{\Xi}X'\left(I_n - W\Xi^{-1}W'\right)\left(W\gamma_0 + e\right). \end{aligned}$$

Then, the bias of $\hat{\beta}_{\Xi}(\Xi)$ is

$$E\left[\Upsilon_{\Xi}X'\left(I_n - W\Xi^{-1}W'\right)W\right]E\left[\gamma_0\right] + E\left[\Upsilon_{\Xi}X'\left(I_n - W\Xi^{-1}W'\right)E\left[e|X,W\right]\right] = 0.$$

□

Proof of Proposition 2. The conditional variance of $\hat{\beta}_\Xi(\Xi)$ is

$$\begin{aligned} var\left(\hat{\beta}_\Xi(\Xi)\,|X,W\right) &= \Upsilon_\Xi X'\left(I_n - W\Xi^{-1}W'\right)\left(W\Gamma W'+\Sigma\right)\left(I_n - W\Xi^{-1}W'\right)' X\Upsilon_\Xi \\ &= \left(X'\Psi_\Xi X\right)^{-1} X'\Psi_\Xi\left(W\Gamma W'+\Sigma\right)\Psi_\Xi X\left(X'\Psi_\Xi X\right)^{-1}, \end{aligned}$$

using Lemma 1, where $\Psi_\Xi = I_n - W\Xi^{-1}W'$. It is well known that this is minimized when the sandwich collapses, which occurs when

$$\Psi_\Xi = \left(W\Gamma W'+\Sigma\right)^{-1},$$

or

$$I_n - W\Xi^{-1}W' = \left(W\Gamma W'+\Sigma\right)^{-1}.$$

Let us find a symmetric Ξ with $m(m+1)/2$ distinct elements that satisfies this system of $n(n+1)/2$ distinct equations. Premultiply by W' and postmultiply by W to get

$$W'W\Xi^{-1}W'W = W'W - W'\left(W\Gamma W'+\Sigma\right)^{-1}W,$$

from where one can express out optimal Ξ:

$$\Xi^* = W'W\left[W'(I_n - \left(W\Gamma W'+\Sigma\right)^{-1})W\right]^{-1}W'W.$$

The corresponding conditional variance is, after collapsing the sandwich,

$$\begin{aligned} var\left(\hat{\beta}_\Xi(\Xi^*)\,|X,W\right) &= \left(X'\Psi_{\Xi^*}X\right)^{-1} \\ &= \left(X'\left(W\Gamma W'+\Sigma\right)^{-1}X\right)^{-1}. \end{aligned}$$

Finally, by the Woodbury matrix identity,

$$\begin{aligned} \Xi^* &= W'W\Big[\left(W'W\right)^{-1} - \left(W'W\right)^{-1}W'(W\left(W'W\right)^{-1}W' - W\Gamma W' - \Sigma)^{-1}W\left(W'W\right)^{-1}\Big]W'W \\ &= W'(I_n - \left(P_W - W\Gamma W' - \Sigma\right)^{-1})W. \end{aligned}$$

□

Proof of Proposition 4. From the proof of Proposition 1, we have

$$\hat{\beta}_\Xi(\Xi^*) = \Upsilon_{\Xi^*}X'\left(I_n - W\Xi^{*-1}W'\right)Y,$$

where $\Upsilon_{\Xi^*} = \left(X'X - X'W\Xi^{*-1}W'X\right)^{-1}$. From the result of Proposition 2, we have $I_n - W\Xi^{*-1}W' = (W\Gamma W'+\Sigma)^{-1}$, hence $\Upsilon_{\Xi^*} = \left(X'\left(W\Gamma W'+\Sigma\right)^{-1}X\right)^{-1}$, and so

$$\begin{aligned} \hat{\beta}_\Xi(\Xi^*) &= \left(X'\left(W\Gamma W'+\Sigma\right)^{-1}X\right)^{-1}X'\left(W\Gamma W'+\Sigma\right)^{-1}Y \\ &= \hat{\beta}_{GLS}. \end{aligned}$$

□

Derivations for Proposition 5. For the least squares estimator,

$$n \cdot var\big(\hat{\beta}_{LS}|X, W\big) = \sigma^2 \left(\frac{X' M_W X}{n}\right)^{-1} \overset{P}{\to} \frac{\sigma^2}{1-\mu},$$

because

$$E\left[\frac{X' M_W X}{n}\right] = \frac{1}{n} E\left[\operatorname{tr}\left(M_W X X'\right)\right] = E\left[\operatorname{tr}\left(M_W E\left[X X'\right]\right)\right] = \frac{1}{n} E\left[\operatorname{tr}\left(M_W\right)\right] = 1 - \mu + o(n^{-1/2}),$$

while the variance asymptotically vanishes as $m, n \to \infty$ (see, for example, Hansen, Hausman and Newey, 2008).

In Design A, for the optimal ridge-out estimator,

$$var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X, W\big) = \sigma^2 \left(X' \left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1} X\right)^{-1}.$$

To determine the limit, let us compute

$$\begin{aligned}
\frac{1}{n} E\left[X' \left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1} X\right] &= \frac{1}{n} E\left[\operatorname{tr}\left(\left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1} E\left[X X'\right]\right)\right] \\
&= \frac{1}{n} E\left[\operatorname{tr}\left(\left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1}\right)\right] \\
&= \frac{1}{n} E\left[\sum_{j=1}^{n} \lambda_j \left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1}\right] \\
&= \frac{1}{n} \sum_{j=1}^{n} E\left[\left(1 + \frac{\omega + o(n^{-1/2})}{n} \lambda_j\left(W W'\right)\right)^{-1}\right].
\end{aligned}$$

Now, $\lambda_j\left(W W'\right) = \lambda_j\left(W' W\right)$ for the m non-zero eigenvalues; the other $n - m$ eigenvalues are equal to zero. Therefore,

$$\begin{aligned}
\frac{1}{n} E\left[X' \left(I_n + \frac{\alpha^2}{m\sigma^2} W W'\right)^{-1} X\right] &= \frac{1}{n} \sum_{j=1}^{m} E\left[\left(1 + \frac{\omega + o(n^{-1/2})}{n} \lambda_j\left(W' W\right)\right)^{-1}\right] \\
&\quad + \frac{1}{n} \sum_{j=m+1}^{n} E\left[\left(1 + \frac{\omega + o(n^{-1/2})}{n} 0\right)^{-1}\right] \\
&= \mu E\big[\big(1 + \omega \lambda(\hat{\Sigma}_W)\big)^{-1}\big] + (1 - \mu) + o(n^{-1/2}),
\end{aligned}$$

where $\hat{\Sigma}_W = n^{-1} W' W$ is the sample variance of W.

Let $a = \left(1-\sqrt{\mu}\right)^2$ and $b = \left(1+\sqrt{\mu}\right)^2$. By the Marchenko-Pastur law (see, for example, Bai and Silverstein 2010, section 3), in the dimension asymptotic limit,

$$\begin{aligned} E\left[\left(1+\omega\lambda\left(\frac{W'W}{n}\right)\right)^{-1}\right] &\to \int_a^b \frac{1}{1+\omega x}\frac{1}{2\pi x\mu}\sqrt{(b-x)(x-a)}dx \\ &= \frac{1}{2\omega\mu}\big(\sqrt{1+\omega^2\left(1-\mu\right)^2+2\omega\left(1+\mu\right)}-1-\omega\left(1-\mu\right)\big). \end{aligned}$$

Therefore,

$$\begin{aligned} E\big[n\cdot var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)\big] &\to \sigma^2\left(\frac{1}{2\omega\mu}\big(\sqrt{1+\omega^2\left(1-\mu\right)^2+2\omega\left(1+\mu\right)}-1-\omega\left(1-\mu\right)\big)+\left(1-\mu\right)\right)^{-1} \\ &= 2\sigma^2\big(\sqrt{\left(\omega^{-1}+\mu-1\right)^2+4\omega^{-1}}-\left(\omega^{-1}+\mu-1\right)\big)^{-1}. \end{aligned}$$

Because $var\big(\hat{\beta}_{LS}|X,W\big)-var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)$ is positive semi-definite, $var\big[n\cdot var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)\big]$ is dominated by $var\big[n\cdot var\big(\hat{\beta}_{LS}|X,W\big)\big]$, which asymptotically vanishes as $m,n\to\infty$. Therefore, $n\cdot var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)$ converges to the above limit.

In Design B, for the optimal ridge-out estimator,

$$var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big) = \sigma^2\left(X'\left(I_n+\frac{\alpha^2}{\sigma^2}P_W\right)^{-1}X\right)^{-1}.$$

To determine the limit, let us compute

$$\begin{aligned} \frac{1}{n}E\left[X'\left(I_n+\frac{\alpha^2}{\sigma^2}P_W\right)^{-1}X\right] &= \frac{1}{n}E\left[\mathrm{tr}\left(\left(I_n+\frac{\alpha^2}{\sigma^2}P_W\right)^{-1}E\left[XX'\right]\right)\right] \\ &= \frac{1}{n}E\left[\sum_{j=1}^{n}\lambda_j\left(I_n+\frac{\alpha^2}{\sigma^2}P_W\right)^{-1}\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}E\left[\left(1+(\mu\omega+o(n^{-1/2}))\lambda_j\left(P_W\right)\right)^{-1}\right]. \end{aligned}$$

Now, $\lambda_j\left(P_W\right)$ are m eigenvalues of unity and $n-m$ eigenvalues of zero. Therefore,

$$\begin{aligned} \frac{1}{n}E\left[X'\left(I_n+\frac{\alpha^2}{\sigma^2}P_W\right)^{-1}X\right] &= \frac{1}{n}\sum_{j=1}^{m}E\left[\left(1+\mu\omega+o(n^{-1/2})\right)^{-1}\right] \\ &\quad+\frac{1}{n}\sum_{j=m+1}^{n}E\left[\left(1+(\mu\omega+o(n^{-1/2}))\cdot 0\right)^{-1}\right] \\ &\to \frac{\mu}{1+\mu\omega}+1-\mu = 1-\frac{\mu^2\omega}{1+\mu\omega}, \end{aligned}$$

Because $var\big(\hat{\beta}_{LS}|X,W\big)-var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)$ is positive semi-definite, $var\big[n\cdot var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)\big]$ is dominated by $var\big[n\cdot var\big(\hat{\beta}_{LS}|X,W\big)\big]$, which asymptotically vanishes as $m,n\to\infty$. Therefore,

$$n\cdot var\big(\hat{\beta}_{\Xi}\left(\Xi^*\right)|X,W\big)\xrightarrow{P}\sigma^2\left(1-\frac{\mu^2\omega}{1+\mu\omega}\right)^{-1}.$$

□

References

Anatolyev, S. (2012): "Inference in regression models with many regressors," *Journal of Econometrics*, 170(2), 368–382.

Anatolyev, S. (2020): "A ridge to homogeneity for linear models," *Journal of Statistical Computation and Simulation*, 90(13), 2455–2472.

Angrist, J., and J. Hahn (2004): "When to control for covariates? Panel asymptotics for estimates of treatment effects," *Review of Economics and Statistics*, 86(1), 58–72.

Bai, Z. and J.W. Silverstein (2010): *Spectral Analysis of Large Dimensional Random Matrices*, Springer-Verlag, New York.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Inference in linear regression models with many covariates and heteroscedasticity," *Journal of the American Statistical Association*, 113(523), 1350–1361.

Dicker, L. (2016): "Ridge regression and asymptotic minimax estimation over spheres of growing dimension," *Bernoulli*, 22(1), 1–37.

Dicker, L. and M. Erdogdu (2017): "Flexible results for quadratic forms with applications to variance components estimation," *Annals of Statistics*, 45(1), 386–414.

Dobriban, E. and S. Wager (2018): "High-dimensional asymptotics of prediction: Ridge regression and classification," *Annals of Statistics.* 46(1), 247–279.

Frisch, R, and F.V. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1(4), 387–401.

Hansen, C., J. Hausman and W.K. Newey (2008): "Estimation with many instrumental variables," *Journal of Business & Economics Statistics*, 26, 398–422.

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

Jiang, J. (1996): "REML estimation: asymptotic behavior and related topics," *Annals of Statistics*, 24(1), 255–286.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-out estimation of variance components," *Econometrica*, 88(5), 1859–1898.

Liu, S. and E. Dobriban (2020): “Ridge regression: structure, cross-validation, and sketching,” *International Conference on Learning Representations (ICLR).*

Lovell, M. (1963): “Seasonal adjustment of economic time series and multiple regression analysis,” *Journal of the American Statistical Association*, 58(304), 993–1010.

Park, J. (2017): “Tolerance intervals from ridge regression in the presence of multicollinearity and high dimension,” *Statistics & Probability Letters*, 121, 128–135.

Robinson, G.K. (1991): “That BLUP is a good thing: the estimation of random effects,” *Statistical Science*, 6(1), 15–32.

van Wieringen, W.N. (2023): “Lecture notes on ridge regression,” arXiv:1509.09169v8 [stat.ME].

van Wieringen, W.N. and C.F.W. Peeters (2016): “Ridge estimation of inverse covariance matrices from high-dimensional data,” *Computational Statistics & Data Analysis*, 103, 284–303.