Supplementary Material to: Sequential learning of regression models by penalized estimation

Wessel N. van Wieringen[1,2,*], **Harald Binder**[3]

[1] Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

[2] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands

[3] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

[*]Corresponding author. Email: w.vanwieringen@amsterdamumc.nl

SM I: Moments of updated ridge regression estimator

The expectation of the nonzero-centered ridge regression estimator is:

$$\begin{aligned}
\mathbb{E}\big[\hat{\boldsymbol{\beta}}(\lambda, \boldsymbol{\beta}_0)\big] &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}[\mathbf{X}^\top\mathbb{E}(\mathbf{Y}) + \mathbb{E}(\lambda\boldsymbol{\beta}_0)] \\
&= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}[\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\mathbb{E}(\boldsymbol{\beta}_0)] \\
&= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}[\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta} - \lambda\boldsymbol{\beta} + \lambda\mathbb{E}(\boldsymbol{\beta}_0)] \\
&= \boldsymbol{\beta} + \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}[\lambda\mathbb{E}(\boldsymbol{\beta}_0) - \boldsymbol{\beta}].
\end{aligned}$$

Similarly, the variance, denoted by the $\mathbb{V}(\cdot)$ operator, of this estimator is:

$$\begin{aligned}
\mathbb{V}\big[\hat{\boldsymbol{\beta}}(\lambda, \boldsymbol{\beta}_0)\big] &= \mathbb{V}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^\top\mathbf{Y} + \lambda\boldsymbol{\beta}_0)] \\
&= \mathbb{V}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{Y}] + \lambda^2\mathbb{V}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\boldsymbol{\beta}_0)] \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1} \\
&\quad +\lambda^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbb{V}[\boldsymbol{\beta}_0)](\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1},
\end{aligned}$$

The expectation of the updated ridge regression estimator is:

$$\begin{aligned}
\mathbb{E}\big[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\big] &= \boldsymbol{\beta} - \lambda_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\boldsymbol{\beta} + \lambda_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\mathbb{E}[\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2})] \\
&= \boldsymbol{\beta} + \lambda_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\boldsymbol{\beta} - \lambda_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\boldsymbol{\beta} \\
&\quad - \lambda_t\lambda_{t-1}\Big\{\prod_{\tau=t-1}^{t}(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big\}\boldsymbol{\beta} \\
&\quad + \Big\{\prod_{\tau=t-1}^{t}[\lambda_\tau(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}]\Big\}\mathbb{E}[\hat{\boldsymbol{\beta}}_{t-2}(\lambda_{t-2}, \hat{\boldsymbol{\beta}}_{t-3})] \\
&= \cdots \\
&= \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h+1}^{t}[\lambda_\tau(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}]^{I_{\{t\geq\tau\}}}\Big\}\boldsymbol{\beta} \\
&\quad - \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h}^{t}[\lambda_\tau(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}]\Big\}\boldsymbol{\beta} + \Big\{\prod_{\tau=1}^{t}[\lambda_\tau(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}]\Big\}\boldsymbol{\beta}_0,
\end{aligned}$$

while its variance is:

$$\begin{aligned}
\mathbb{V}\big[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\big] &= \sigma^2(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\mathbf{X}_t^\top\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1} \\
&\quad +\lambda_t^2(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\mathbb{V}[\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2})](\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1} \\
&= \sigma^2(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1}\mathbf{X}_t^\top\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t + \lambda_t\mathbf{I}_{pp})^{-1} \\
&\quad +\sigma^2\lambda_t^2\Big[\prod_{\tau=t-1}^{t}(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big]\mathbf{X}_{t-1}^\top\mathbf{X}_{t-1}\Big[\prod_{\tau=t-1}^{t}(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big] \\
&\quad +\Big[\prod_{\tau=t-1}^{t}(\lambda_{s,\tau}\lambda_{\ell,\tau}\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_{s,\tau}\mathbf{I}_{pp})^{-1}\Big]\mathbb{V}[\hat{\boldsymbol{\beta}}_{t-2}(\lambda_{s,t-2}, \hat{\boldsymbol{\beta}}_{t-3})]\Big[\prod_{\tau=t-1}^{t}(\lambda_\tau\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big] \\
&= \cdots \\
&= \sum_{t_h=1}^{t}\sigma^2\Big[\prod_{\tau=t_h+1}^{t}(\lambda_\tau^2)^{I_{\{t\geq\tau\}}}\Big]\Big[\prod_{\tau=t_h}^{t}(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big]\mathbf{X}_{t_h}^\top\mathbf{X}_{t_h}\Big[\prod_{\tau=t_h}^{t}(\mathbf{X}_\tau^\top\mathbf{X}_\tau + \lambda_\tau\mathbf{I}_{pp})^{-1}\Big].
\end{aligned}$$

Assume $\lambda_t = \lambda$ and an orthogonal design matrix, i.e. $\mathbf{X}_t = \mathbf{I}_{pp}$. Then, the moments of iterative ridge regression estimator simplify to:

$$\begin{aligned}
\mathbb{E}\big[\hat{\boldsymbol{\beta}}_t(\lambda_t)\big] &= \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h+1}^{t}[\lambda(\mathbf{I}_{pp}+\lambda\mathbf{I}_{pp})^{-1}]^{I_{\{t\geq\tau\}}}\Big\}\boldsymbol{\beta} \\
&\quad - \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h}^{t}[\lambda(\mathbf{I}_{pp}+\lambda\mathbf{I}_{pp})^{-1}]\Big\}\boldsymbol{\beta} + \Big\{\prod_{\tau=1}^{t}[\lambda(\mathbf{I}_{pp}+\lambda\mathbf{I}_{pp})^{-1}]\Big\}\boldsymbol{\beta}_0 \\
&= \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h+1}^{t}[\lambda(1+\lambda)^{-1}]^{I_{\{t\geq\tau\}}}\Big\}\boldsymbol{\beta} - \sum_{t_h=1}^{t}\Big\{\prod_{\tau=t_h}^{t}[\lambda(1+\lambda)^{-1}]\Big\}\boldsymbol{\beta} + \Big\{\prod_{\tau=1}^{t}[\lambda(1+\lambda)^{-1}]\Big\}\boldsymbol{\beta}_0 \\
&= \lambda^{-1}(1+\lambda)\sum_{t_h=1}^{t}[\lambda^t(1+\lambda)^{-t}]\boldsymbol{\beta} - \sum_{t_h=1}^{t}[\lambda^t(1+\lambda)^{-t}]\boldsymbol{\beta} + \lambda^t(1+\lambda)^{-t}\boldsymbol{\beta}_0 \\
&= \lambda^{-1}\sum_{t_h=1}^{t}[\lambda^t(1+\lambda)^{-t}]\boldsymbol{\beta} + \lambda^t(1+\lambda)^{-t}\boldsymbol{\beta}_0 \\
&= \lambda^{-1}\frac{\lambda(1+\lambda)^{-1}}{1-\lambda(1+\lambda)^{-1}}[1-\lambda^t(1+\lambda)^{-t}]\boldsymbol{\beta} + \lambda^t(1+\lambda)^{-t}\boldsymbol{\beta}_0 \\
&= [1-\lambda^t(1+\lambda)^{-t}]\boldsymbol{\beta} + \lambda^t(1+\lambda)^{-t}\boldsymbol{\beta}_0 \\
&= \boldsymbol{\beta} + \lambda^t(1+\lambda)^{-t}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}),
\end{aligned}$$

and its variance

$$\begin{aligned}
\mathrm{Var}\big[\hat{\boldsymbol{\beta}}_t(\lambda_t)\big] &= \sigma_\varepsilon^2\sum_{t_h=1}^{t}\Big[\prod_{\tau=t_h+1}^{t}(\lambda^2)^{I_{\{t\geq\tau\}}}\Big]\Big[\prod_{\tau=t_h}^{t}(\mathbf{I}_{pp}+\lambda\mathbf{I}_{pp})^{-1}\Big]\mathbf{I}_{pp}\Big[\prod_{\tau=t_h}^{t}(\mathbf{I}_{pp}+\lambda\mathbf{I}_{pp})^{-1}\Big] \\
&= \sigma_\varepsilon^2\sum_{t_h=1}^{t}\Big[\lambda^{-2}\prod_{\tau=t_h}^{t}\lambda^2\Big]\Big[\prod_{\tau=t_h}^{t}(1+\lambda)^{-1}\Big]^2\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2\Big[\lambda^{-2}\sum_{t_h=1}^{t}\lambda^{2(t-t_h+1)}(1+\lambda)^{-2(t-t_h+1)}\Big]\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2\Big\{\lambda^{-2}\sum_{t_h=1}^{t}[\lambda^2(1+\lambda)^{-2}]^{t-t_h+1}\Big\}\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2\Big\{\lambda^{-2}\sum_{t_h=1}^{t}[\lambda^2(1+\lambda)^{-2}]^{t_h}\Big\}\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2\lambda^{-2}\Big[\frac{\lambda^2(1+\lambda)^{-2}}{1-\lambda^2(1+\lambda)^{-2}} - \lambda^{2t}(1+\lambda)^{-2t}\frac{\lambda^2(1+\lambda)^{-2}}{1-\lambda^2(1+\lambda)^{-2}}\Big]\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2\frac{(1+\lambda)^{-2}}{1-\lambda^2(1+\lambda)^{-2}}\big[1-\lambda^{2t}(1+\lambda)^{-2t}\big]\mathbf{I}_{pp} \\
&= \sigma_\varepsilon^2(1+2\lambda)^{-1}\big[1-\lambda^{2t}(1+\lambda)^{-2t}\big]\mathbf{I}_{pp}
\end{aligned}$$

SM II: Proofs

Theorem 1. *(Asymptotic unbiasedness)*
Assume the existence of an infinite sequence of studies into the linear relationship between a continuous response and a set of covariates. The data from these studies, $\{\mathbf{X}_t, \mathbf{Y}_t\}_{t=1}^{\infty}$, are used to fit the linear regression model by means of the updated ridge linear regression estimator, which yields the sequence of estimators $\{\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^{\infty}$ which is initiated by an arbitrary, nonrandom $\boldsymbol{\beta}_0$. Furthermore, $T \in \mathbb{N}$ be sufficiently large and $\mathbf{X}_{new}$ be the design matrix with covariate information on novel samples for which a prediction is needed. Then,

- $\lim_{t\to\infty} \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+1}] = \boldsymbol{\beta} + \mathbf{u}$ *for some* $\mathbf{u} \in \cap_{t=T}^{\infty} null(\mathbf{X}_t)$*, where* $null(\mathbf{X}_t)$ *denotes the null space of the linear map induced by* $\mathbf{X}_t$*. If* $\cap_{t=T}^{\infty} null(\mathbf{X}_t) = \mathbf{0}_p$*, then* $\mathbf{u} = \mathbf{0}_p$*.*
- $\lim_{t\to\infty} \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\mathbf{X}_{new}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+1}] = \mathbf{X}_{new}\boldsymbol{\beta}$*.*

Proof.
Stationarity of the estimator
Theorem 8.2.14 of [4] lays out the conditions for the existence of and convergence to a stationary distribution of a discrete time, time-homogeneous Markov process with a continuous state space. To show the stationarity of the Markov process of the ridge updated linear regression estimator these conditions need to be verified. The conditions comprise *i)* the irreducibility of the process, i.e. it should satisfy the mixing condition, *ii)* the geometric drift of the process to the center, and *iii)* the uniform integrability of the sequence of the process' marginal densities. Conditions *i)* and *ii)* are verified next, as third one follows from a general argument laid out in [4] and applicable here.

The sequence $\{\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t)\}_{t=1}^{\infty}$ has a stationary/invariant distribution. The irreducibility/mixing condition requires to verify that with positive probability any $\boldsymbol{\beta}'$ in the state space $\mathcal{S}$, a compact subset of $\mathbb{R}^p$ (compactness is due to the penalization), is reachable (after finite time) from any $\boldsymbol{\beta}'' \in \mathcal{S}$. The unpenalized linear regression estimator is unconstrained and may – with positive probability – assume any value in $\mathcal{S}$. The penalty restricts the estimator to a circular domain around its center that is formed by the target, here the previous updated regression estimate. Hence, it is not the nonzero center but only the value of penalty parameter that limits the state space of the estimator. But as the penalty parameter is chosen in a data-driven manner, it is itself random. Put differently, $\lambda_{t+1} > 0$ follows some distribution which assigns a positive probability to any value on $\mathbb{R}_{>0}$. Hence, at any t, the parameter constraint can (with positive probability) be arbitrarily large, thus allowing any value (of $\mathcal{S}$). However, the ridge regression estimator at time $t+1$ is of the form:

$$\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \quad = \quad (\mathbf{X}_{t+1}^\top \mathbf{X}_{t+1} + \lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top \mathbf{Y}_{t+1} + \lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})].$$

From which it clear that its values are constrained to the (affine) subspace spanned by the rows of $\mathbf{X}_{t+1}$. But $\boldsymbol{\beta}'$ needs only to be reachable from $\boldsymbol{\beta}''$ in finite time. Two scenario's are possible: $\cap_{t=T}^{T+T'} \text{null}(\mathbf{X}_t) = \mathbf{0}_p$ or $\cap_{t=T}^{T+T'} \text{null}(\mathbf{X}_t) \neq \mathbf{0}_p$ for $T, T' \in \mathbb{N}$ and T' large enough. The former scenario warrants that indeed any value in $\mathcal{S}$ is reachable after finite time. The latter scenario implies that the state space is reducible to $\mathcal{S} \setminus \cap_{t=T}^{T+T'} \text{null}(\mathbf{X}_t)$, which we denote by $\mathcal{S}'$. In the latter case sufficient mixing does happen within the reduced state space $\mathcal{S}'$.

To assess the geometric drift to the center, we derive the following inequality:

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\|_F &= \|(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F \\
&\leq \|(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\quad + \|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F \\
&\leq \|(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\quad + \|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\|_2\,\|\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F,
\end{aligned}$$

in which the triangular inequality is invoked twice. Now let $\mathbf{X}_{t+1}=\mathbf{U}_{t+1}\mathbf{D}_{t+1}\mathbf{V}_{t+1}^\top$ be the singular value decomposition of $\mathbf{X}_{t+1}$ with $\mathbf{U}_{t+1}$ and $\mathbf{V}_{t+1}$ the matrices with the left and right (respectively) singular vectors as columns and $\mathbf{D}_{t+1}$ the diagonal matrix with the singular values on its diagonal. The eigenvalues of $\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}$ are then to be found on the diagonal of $\lambda_{t+1}(\mathbf{D}_{t+1}^2+\lambda_{t+1}\mathbf{I}_{pp})^{-1}$. These diagonal elements equal one when the corresponding singular value is zero and inside the unit interval otherwise. Consequently, $\|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\|_2<1$. There thus exits $C>0$ and $\alpha\in(0,1)$ such that $\|\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\|_F\leq C+\alpha\|\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F$. From which we conclude the tightness of the sequence. The defined Markov chain now has, by Theorem 8.2.14 of [4], a stationary distribution, which is reached after some finite T.

Stationarity of the linear predictor

Stationarity of the linear predictor is proven along the same line as that of the estimator. For the stationarity we now only show the mixing condition and the geometric drift to the center (as the rest is similar) of the process. It is then left to identify the limit. Note that the irreducibility of the process of updated linear predictors follows from that of the ridge updated linear regression estimators, as the multiplication of these estimators by $\mathbf{X}_{\text{new}}$ maps this process onto $\mathbb{R}$, which is a surjective function.

To assess the geometric drift to the center, we derive the following inequality:

$$\begin{aligned}
&\|\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\|_F \\
&\quad= \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F \\
&\quad\leq \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\qquad + \|\lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F \\
&\quad= \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\qquad + \left(\text{tr}\{\mathbf{X}_{\text{new}}\mathbf{M}_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]^\top\mathbf{M}_{t+1}\mathbf{X}_{\text{new}}^\top\}\right)^{1/2} \\
&\quad\leq \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\qquad + \left(\|\mathbf{M}_{t+1}\|_\infty\text{tr}\{\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]^\top\mathbf{M}_{t+1}\mathbf{X}_{\text{new}}^\top\mathbf{X}_{\text{new}}\}\right)^{1/2} \\
&\quad\leq \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F \\
&\qquad + \left(\|\mathbf{M}_{t+1}\|_\infty^2\text{tr}\{\mathbf{X}_{\text{new}}^\top\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]^\top\}\right)^{1/2} \\
&\quad\leq \|\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}\|_F + \|\mathbf{M}_{t+1}\|_\infty\|\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\|_F,
\end{aligned}$$

where $\mathbf{M}_{t+1}=[\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}]$ and we have used the triangular inequality, the Hölder inequality for the p-Schatten norm, the cyclic property of the trace, and the fact that $\mathbf{M}_{t+1}$ has eigenvalues only in the unit interval. There thus exits $C>0$ and $\alpha\in(0,1)$ such that $\|\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\|_F\leq C+\alpha\|\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\|_F$. The sequence of updated linear predictors is thus tight. From Theorem 8.2.14 of [4] we now conclude that the defined Markov chain has a stationary distribution, which is reached after some finite T.

Asymptotic expectation of the estimator

The defined Markov chain now has, by Theorem 8.2.14 of [4], a stationary distribution, which is reached after some finite T. Once stationarity has been reached: $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}] = \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]$. Use this is

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]\\
&\quad= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}\{(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}\}\\
&\quad= (\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\{\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}\boldsymbol{\beta}+\lambda_{t+1}\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}\}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]\\
&\quad= (\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}].
\end{aligned}$$

To arrive at:

$$(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}[\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})-\boldsymbol{\beta}\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}] \;=\; \mathbf{0}_p. \tag{1}$$

When $\mathbf{X}_{t+1}$ is of full column rank, this is implies $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]=\boldsymbol{\beta}$. If not, $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}] = \boldsymbol{\beta}+\mathbf{u}$ with $\mathbf{u}\in\text{null}(\mathbf{X}_{t+1})$. But as $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]=\boldsymbol{\beta}_\infty$ for all t larger than some T for which stationary has been reached, Equation (1) then implies that $\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}(\boldsymbol{\beta}_\infty-\boldsymbol{\beta})=\mathbf{0}_p$ for all $t\geq T$. Hence, if $\cap_{t=T}^\infty\text{null}(\mathbf{X}_t)=\mathbf{0}_p$, we conclude that updated ridge regression estimator is asymptotically unbiased: $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]=\boldsymbol{\beta}$ as $t\to\infty$.

Asymptotic expectation of the linear predictor

The result is proven below for a single sample. Due to the usual assumption of independence of samples, it applies to larger sample sizes.

For the asymptotic unbiasedness, recall from above that, for $t>T$ we have that $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t}] = \boldsymbol{\beta}+\mathbf{u}$ with $u\in\cap_{t=T}^\infty\text{null}(\mathbf{X}_t)$. Hence, $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}] = \mathbf{X}_{\text{new}}\boldsymbol{\beta}+\mathbf{X}_{\text{new}}\mathbf{u}$ with $\mathbf{u}\in\cap_{t=T}^\infty\text{null}(\mathbf{X}_t)$. It thus rests to show that $\mathbf{X}_{\text{new}}\mathbf{u}=0$. This requires stationarity of the sequence of predictors $\{\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\}_{t=1}^\infty$, which we have shown above.

We are now ready to show that $\mathbf{X}_{\text{new}}\mathbf{u}=0$. Note that, once stationarity has been reached: $\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}] = \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t}]$. Use this in the following:

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\boldsymbol{\beta}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}]\\
&\quad= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}\{\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\,|\,\{\lambda_\tau\}_{\tau=1}^{t+1}\}\\
&\quad= \mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}\boldsymbol{\beta}\\
&\qquad+\lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t}]\\
&\quad= \mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp}-\lambda_{t+1}\mathbf{I}_{pp})\boldsymbol{\beta}\\
&\qquad+\lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t}]\\
&\quad= \mathbf{X}_{\text{new}}\boldsymbol{\beta}+\lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^{t}]-\boldsymbol{\beta}\}.
\end{aligned}$$

Reformulated and using stationarity:

$$\begin{aligned}
&\mathbf{X}_{\text{new}}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\\
&\quad = \quad \lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}.
\end{aligned}$$

Now use the 'trace trick' and the Hölder inequality for the matrix Schatten norm to obtain the following inequality:

$$\begin{aligned}
&|\text{tr}\big(\mathbf{X}_{\text{new}}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\big)|\\
&\quad = \quad |\text{tr}\big(\lambda_{t+1}\mathbf{X}_{\text{new}}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\big)|\\
&\quad \leq \quad \|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\|_\infty|\text{tr}\big(\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\mathbf{X}_{\text{new}}\big)|\\
&\quad = \quad \|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\|_\infty|\text{tr}\big(\mathbf{X}_{\text{new}}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\big)|.
\end{aligned}$$

But $\|\lambda_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\|_\infty \in (0,1)$ for positive λ_{t+1} and we must thus have that $\text{tr}\big(\mathbf{X}_{\text{new}}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}=0$. Or, equivalently, that

$$\text{tr}\big(\mathbf{X}_{\text{new}}\{\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]-\boldsymbol{\beta}\}\big) \quad = \quad \text{tr}(\mathbf{X}_{\text{new}}\mathbf{u}) \quad = \quad \mathbf{X}_{\text{new}}\mathbf{u} \quad = \quad 0.$$

The updated ridge linear regression predictor is thus asymptotically unbiased. □

Theorem 2. *(Consistency of the updated ridge linear regression estimator and predictor) Assume an infinite sequence of studies into the linear relationship between a continuous response and a set of covariates. The data from these studies, $\{\mathbf{X}_t,\mathbf{Y}_t\}_{t=1}^\infty$, are used to fit the linear regression model by means of the updated ridge linear regression estimator, which yields the sequence of estimators $\{\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^\infty$ which is initiated by an arbitrary, nonrandom $\boldsymbol{\beta}_0$. Let $T\in\mathbb{N}$ be sufficiently large and $\cap_{t=T}^\infty null(\mathbf{X}_t)=\mathbf{0}_p$. Assume that the penalty parameter sequence $\{\lambda_t\}_{t=1}^\infty$ is chosen such that $\lim_{t\to\infty}\sigma_\varepsilon^2\,p\,d_1^2(\mathbf{X}_t)\,\lambda_t^{-2}=0$ with $d_1(\mathbf{X}_t)$ the largest singular value of $\mathbf{X}_t$. Then, for every $c>0$:*

$$\lim_{t\to\infty}P[\|\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})-\boldsymbol{\beta}\|\geq c\,|\,\{\lambda_\tau\}_{\tau=1}^t]\to 0,$$
$$\lim_{t\to\infty}P[\|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})-\mathbf{X}_{new}\boldsymbol{\beta}\|\geq c\,|\,\{\lambda_\tau\}_{\tau=1}^t]\to 0.$$

Proof. To prove convergence in probability for the updated ridge regression estimator, we assume, without loss of generality that, that the sequence $\{\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\}_{k=1}^\infty$ is initiated by the stationary density (if not, it will become stationary after finite time, cf. the proof of Theorem 1). By the condition on the intersection of the null spaces of the design matrices and Theorem 1 the updated ridge regression estimator is then unbiased. This leaves to prove that its variance vanishes as t tends to infinity. Theorem 8.2 of [2], by Chebyshev's inequality then, warrants the convergence in probability of the estimator.

To prove that the variance of $\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})$ vanishes as $t\to\infty$, note that

$$\begin{aligned}
&\text{tr}\{\text{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}\\
&\quad = \quad \text{tr}\{\sigma_\varepsilon^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\}\\
&\qquad +\text{tr}\{\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\text{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t](\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}\}\\
&\quad = \quad \sigma_\varepsilon^2\sum\nolimits_{j=1}^p(d_{t+1,j}^2+\lambda_{t+1})^{-2}d_j^2(\mathbf{X}_{t+1})\\
&\qquad +\text{tr}\{\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-2}\text{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}\\
&\quad \leq \quad \sigma_\varepsilon^2 p\lambda_{t+1}^{-2}d_1^2(\mathbf{X}_{t+1})+\text{tr}\{\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-2}\text{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}.
\end{aligned}$$

Furthermore, as all eigenvalues of $\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-2}$ are in the interval $(0,1]$ with at least one smaller than one, we have:

$$\begin{aligned}&\mathrm{tr}\{\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-2}\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}\\ &\quad<\quad \mathrm{tr}\{\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\},\end{aligned}$$

in which we have used von Neumann's trace inequality. In particular, there exists a $\alpha_{t+1}\in(0,1)$ such that

$$\begin{aligned}&\mathrm{tr}\{\lambda_{t+1}^2(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-2}\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}\\ &\quad=\quad \alpha_{t+1}\mathrm{tr}\{\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}.\end{aligned}$$

Put together,

$$\begin{aligned}&\mathrm{tr}\{\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}\\ &\quad<\quad \sigma_\varepsilon^2 p d_1^2(\mathbf{X}_{t+1})\lambda_{t+1}^{-2}+\alpha_{t+1}\mathrm{tr}\{\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\}.\end{aligned}$$

Hence, if the penalty parameter sequence $\{\lambda_t\}_{t=1}^\infty$ is chosen such that $\lim_{t\to\infty}\sigma_\varepsilon^2\,p\,d_1^2(\mathbf{X}_{t+1})\,\lambda_{t+1}^{-2}=0$, the variance of the estimator vanishes as t increases.

The consistency of the updated ridge linear predictor is, by the Continuous Mapping Theorem (Theorem 2.3, [5]), a direct consequence of the consistency of the updated ridge regression estimator. □

Theorem 3. *(Asymptotics of the updated ridge logistic regression estimator)*
Assume an infinite sequence of studies into the generalized linear relationship between a binary response and a set of covariates. The data from these studies, $\{\mathbf{X}_t,\mathbf{Y}_t\}_{t=1}^\infty$, are used to fit the logistic regression model by means of the updated ridge logistic regression estimator, which yields the sequence of estimators $\{\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})\}_{t=1}^\infty$ which is initiated by an arbitrary, nonrandom $\boldsymbol{\beta}_0$. Let $T\in\mathbb{N}$ be sufficiently large. Then:

- $\lim_{t\to\infty}\mathbb{E}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)]=\boldsymbol{\beta}+\mathbf{u}$ *for some* $\mathbf{u}\in\cap_{t=T}^\infty null(\mathbf{X}_t)$. *If* $\cap_{t=T}^\infty null(\mathbf{X}_t)=\mathbf{0}_p$, *then* $\mathbf{u}=\mathbf{0}_p$.
- *if* $\cap_{t=T}^\infty null(\mathbf{X}_t)=\mathbf{0}_p$ *and* $\{\lambda_t\}_{t=1}^\infty$ *such that* $\lim_{t\to\infty}2p^{1/2}\,|d_1(\mathbf{X}_t)|\,\lambda_t=0$ *with* $d_1(\mathbf{X}_t)$ *the largest singular value of* $\mathbf{X}_t$, *for every* $c>0$*:*

$$\lim_{t\to\infty}P[\|\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})-\boldsymbol{\beta}\|\geq c\,|\,\{\lambda_\tau\}_{\tau=1}^t]\to 0,$$
$$\lim_{t\to\infty}P[\|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})-\mathbf{X}_{new}\boldsymbol{\beta}\|\geq c\,|\,\{\lambda_\tau\}_{\tau=1}^t]\to 0.$$

Proof. ***Asymptotic unbiasedness***
The proof requires the existence of a stationary density of the updating process of ridge estimators, from which the unbiasedness follows by use of the estimating equation of the ridge regression estimator.

The argument for the existence of a stationary distribution can be borrowed from the proof of Theorem 1. Hereto now that the updated ridge logistic regression estimator can, after rescaling, $\tilde{\mathbf{X}}_{t+1}=\mathbf{X}_{t+1}\mathbf{W}_{t+1}^{1/2}$ and $\tilde{\mathbf{Z}}_{t+1}=\mathbf{W}_{t+1}^{1/2}\mathbf{Z}_{t+1}$, be written as:

$$\hat{\boldsymbol{\beta}}_{t+1}[\lambda_{t+1},\hat{\boldsymbol{\beta}}_t]\quad=\quad[\tilde{\mathbf{X}}_{t+1}^\top\tilde{\mathbf{X}}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp}]^{-1}[\tilde{\mathbf{X}}_{t+1}^\top\tilde{\mathbf{Z}}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})],\tag{2}$$

in which one recognizes the form of the updated ridge linear regression estimator. Hence, the arguments employed in the proof of Theorem 1 apply here. Similarly, the arguments in Theorem

1 can also be used to show the stationarity of the sequence of updated ridge logistic regression predictors.

The defined Markov chain now has, by Theorem 8.2.14 of [4], a stationary distribution, which is reached after some finite T. Once stationarity has been reached all estimators $\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1})$ share the same distribution (marginally). Hence,

$$\begin{aligned}\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}] &= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+1}] \qquad = \ldots \\ &= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+t_0}} \hat{\boldsymbol{\beta}}_{t+t_0}(\lambda_{t+t_0}, \hat{\boldsymbol{\beta}}_{t+t_0-1}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+t_0}],\end{aligned}$$

but also

$$\begin{aligned}\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\vec{\mathbf{g}}^{-1}(\mathbf{X}_{t_1}; \hat{\boldsymbol{\beta}}_t) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}] &= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+1}}[\vec{\mathbf{g}}^{-1}(\mathbf{X}_{t_1}; \hat{\boldsymbol{\beta}}_{t+1}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+1}] \quad = \quad \ldots \\ &= \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+t_0}}[\vec{\mathbf{g}}^{-1}(\mathbf{X}_{t_1}; \hat{\boldsymbol{\beta}}_{t+t_0}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t+t_0}]\end{aligned}$$

for all $t_0, t_1 \in \mathbb{N}$. Applied to the estimating equation and aggregated over the t_0:

$$\begin{pmatrix} \mathbf{X}_{t+1} \\ \mathbf{X}_{t+2} \\ \ldots \\ \mathbf{X}_{t+\tau} \end{pmatrix}^\top \left[\mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+t_0}} \left(\begin{pmatrix} \mathbf{Y}_{t+1} \\ \mathbf{Y}_{t+2} \\ \ldots \\ \mathbf{Y}_{t+t_0} \end{pmatrix} \,\middle|\, \{\lambda_\tau\}_{\tau=1}^{t+t_0} \right) - \mathbb{E}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t+t_0}} \left(\begin{pmatrix} \vec{\mathbf{g}}^{-1}(\mathbf{X}_{t+1}; \hat{\boldsymbol{\beta}}_{t+t_0}) \\ \vec{\mathbf{g}}^{-1}(\mathbf{X}_{t+2}; \hat{\boldsymbol{\beta}}_{t+t_0}) \\ \ldots \\ \vec{\mathbf{g}}^{-1}(\mathbf{X}_{t+t_0}; \hat{\boldsymbol{\beta}}_{t+t_0}) \end{pmatrix} \,\middle|\, \{\lambda_\tau\}_{\tau=1}^{t+t_0} \right) \right] = \mathbf{0}.$$

As the elements of the $\mathbf{Y}_t$'s are binomially distributed, their (conditionally) probabilities are consistently estimated by $\lim_{\tau\to\infty} \vec{\mathbf{g}}(\mathbf{X}_t \hat{\boldsymbol{\beta}}_{t+\tau})$. The Continuous Mapping theorem (Theorem 2.3 of [5]) then warrants that $\mathbf{X}_t \hat{\boldsymbol{\beta}}_{t+\tau}$ is a consistent estimator of $\mathbf{X}_t \boldsymbol{\beta}$. In particular, $\mathbf{X}_t$ may be replaced by any $\mathbf{X}_{\text{new}}$. Hence, the updated ridge logistic regression predictor is asymptotically unbiased.

Consistency

To prove convergence in probability for the updated ridge regression estimator, we assume, without loss of generality that, that the sequence $\{\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t)\}_{k=1}^{\infty}$ is initiated by the stationary density (if not, it will become stationary after finite time, see above). By the condition on the intersection of the null spaces of the design matrices and the first part of the theorem, the updated ridge logistic regression estimator is unbiased. This leaves to prove that its variance vanishes as t tends to infinity. Theorem 8.2 of [2], by Chebyshev's inequality then, warrants the convergence in probability of the estimator.

To show that the variance of the updated ridge regresssion estimator vanishes, we study the sum of the variances of the last two updated ridge regression estimators. This sum can be expressed as:

$$\begin{aligned}&\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}]\} + \mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}] \\ &= \tfrac{1}{2}\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) - \hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}] \\ &\quad + \tfrac{1}{2}\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) + \hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1}, \hat{\boldsymbol{\beta}}_{t-2}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}]\end{aligned}$$

For the first summand on the right-hand side of the equality sign, the estimating equation gives:

$$\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) - \hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) = \lambda_{t+1}^{-1} \mathbf{X}_{t+1}^\top \{\mathbf{Y}_{t+1} - \vec{\mathbf{g}}^{-1}[\mathbf{X}_{t+1}; \hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t)]\}.$$

We then obtain the bound:

$$\begin{aligned}&\mathrm{tr}\{\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t) - \hat{\boldsymbol{\beta}}_t(\lambda_t, \hat{\boldsymbol{\beta}}_{t-1}) \,|\, \{\lambda_\tau\}_{\tau=1}^{t}]\} \\ &= \lambda_{t+1}^{-2} \mathrm{tr}\big(\mathbf{X}_{t+1}\mathbf{X}_{t+1}^\top \mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^{t}}\{\mathbf{Y}_{t+1} - \vec{\mathbf{g}}^{-1}[\mathbf{X}_{t+1}; \hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1}, \hat{\boldsymbol{\beta}}_t)] \,|\, \{\lambda_\tau\}_{\tau=1}^{t}\}\big) \\ &\leq \lambda_{t+1}^{-2}\, p\, d_1^2(\mathbf{X}_{t+1})\end{aligned}$$

For the other term, the estimating equation also gives:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})+\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2}) &= 2\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})+\lambda_t^{-1}\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\\ &= 2\lambda_{t-1}^{-1}\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\\ &\quad+\lambda_t^{-1}\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\end{aligned}$$

Insert this identity in the variance of t-th updated ridge logistic regression estimator and obtain:

$$\begin{aligned}&\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})+\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\\ &= \mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\Big(2\lambda_{t-1}^{-1}\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\\ &\qquad+\lambda_t^{-1}\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,|\,\{\lambda_\tau\}_{\tau=1}^t\Big)\\ &= \mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\Big(2\lambda_{t-1}^{-1}\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\,|\,\{\lambda_\tau\}_{\tau=1}^t\Big)\\ &\qquad+\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\Big(\lambda_t^{-1}\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,|\,\{\lambda_\tau\}_{\tau=1}^t\Big)\\ &\quad+2\mathrm{Cov}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\Big(2\lambda_{t-1}^{-1}\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\},\\ &\qquad\lambda_t^{-1}\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,|\,\{\lambda_\tau\}_{\tau=1}^t\Big).\end{aligned}$$

Using the same argumentation as above, we obtain:

$$\begin{aligned}&\mathrm{tr}\Big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})+\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\Big]\\ &\leq \lambda_t^{-2}\,p\,d_1^2(\mathbf{X}_t)+4\lambda_{t-1}^{-2}\,p\,d_1^2(\mathbf{X}_{t-1})\\ &\quad+4\lambda_{t-1}^{-1}\lambda_t^{-1}\mathrm{tr}\Big[\mathrm{Cov}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\Big(\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\},\\ &\qquad\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,|\,\{\lambda_\tau\}_{\tau=1}^t\Big)\Big].\end{aligned}$$

It rests to bound the covariance terms in the expression above. That can be done as follows:

$$\begin{aligned}&\mathrm{tr}\big\{\mathrm{Cov}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\},\\ &\qquad\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big\}\\ &\leq \sum_{j=1}^p\big\{\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big]_{jj}\\ &\qquad\times\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big]_{jj}\big\}^{1/2}\\ &\leq \big\{\mathrm{tr}\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{X}_t^\top\{\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\}\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big]\big\}^{1/2}\\ &\quad\big\{\mathrm{tr}\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{X}_{t-1}^\top\{\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\}\,\big|\,\{\lambda_{\tau'}\}_{\tau'=1}^t\big)\big]\big\}^{1/2}\\ &= \big\{\mathrm{tr}\big[\mathbf{X}_\tau\mathbf{X}_t^\top\mathrm{Var}_{\{\mathbf{Y}_t\}_{\tau=1}^t}\big(\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big]\big\}^{1/2}\\ &\quad\big\{\mathrm{tr}\big[\mathbf{X}_{t-1}\mathbf{X}_{t-1}^\top\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\,\big|\,\{\lambda_{\tau'}\}_{\tau'=1}^t\big)\big]\big\}^{1/2}\\ &\leq |d_1(\mathbf{X}_\tau)|\,|d_1(\mathbf{X}_{\tau'})|\big\{\mathrm{tr}\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{Y}_t-\vec{\mathbf{g}}^{-1}[\mathbf{X}_t;\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})]\,\big|\,\{\lambda_\tau\}_{\tau=1}^t\big)\big]\big\}^{1/2}\\ &\quad\big\{\mathrm{tr}\big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}\big(\mathbf{Y}_{t-1}-\vec{\mathbf{g}}^{-1}[\mathbf{X}_{t-1};\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})]\,\big|\,\{\lambda_{\tau'}\}_{\tau'=1}^t\big)\big]\big\}^{1/2}\\ &\leq |d_1(\mathbf{X}_t)|\,|d_1(\mathbf{X}_{t-1})|\,p^{1/2}\,p^{1/2},\end{aligned}$$

where have used the Cauch-Schwarz inequality and well-known trace inequalities for the product of two nonnegative definite matrices. Put together we then have:

$$\begin{aligned}
&\mathrm{tr}\Big[\mathrm{Var}_{\{\mathbf{Y}_\tau\}_{\tau=1}^t}[\hat{\boldsymbol{\beta}}_t(\lambda_t,\hat{\boldsymbol{\beta}}_{t-1})+\hat{\boldsymbol{\beta}}_{t-1}(\lambda_{t-1},\hat{\boldsymbol{\beta}}_{t-2})\,|\,\{\lambda_\tau\}_{\tau=1}^t]\Big] \\
&\quad\leq\quad \lambda_t^{-2}\,p\,d_1^2(\mathbf{X}_t)+\lambda_{t-1}^{-2}\,p\,d_1^2(\mathbf{X}_{t-1})+4\lambda_{t-1}^{-1}\lambda_t^{-1}|d_1(\mathbf{X}_t)|\,|d_1(\mathbf{X}_{t-1})|p.
\end{aligned}$$

The right-hand side vanishes if $\lim_{t\to\infty} 2p\,|d_1(\mathbf{X}_t)|\,\lambda_t = 0$. □

Theorem 4. *(Mean squared error of mixed vs. updated estimator)*
Let $\mathbf{X}_t$ be orthonormal and $\lambda_t > \sigma_\varepsilon(\sigma_\varepsilon^2+\sigma_\gamma^2)^{-1/2}2^{t/2}T^{1/2}$ for $1\leq t\leq T$. Then, when initiated by $\hat{\boldsymbol{\beta}}_1(\lambda_1)=\hat{\boldsymbol{\beta}}_1^{(me)}$, the updated ridge regression estimator outperforms (in the mean squared error sense), the maximum likelihood estimator of the mixed model's fixed effects parameter: $MSE[\hat{\boldsymbol{\beta}}_T(\lambda_T)] < MSE[\hat{\boldsymbol{\beta}}_T^{(me)}]$.

Proof. The proof derives the mean squared error of both estimators under the orthonormality assumption. Subsequently, using the particulars of penalty parameter scheme, the claimed inequality is shown.

First the mean squared error of the maximum likelihood estimator of the mixed model's fixed effects parameter $\boldsymbol{\beta}$ is obtained. Hereto rewrite the estimator using the Woodbury identity (Theorem 18.2.8, [1]) to:

$$(\xi\mathbb{Z}\mathbb{Z}^\top+\mathbf{I}_{T\tilde{n}_T,T\tilde{n}_T})^{-1} \quad=\quad \mathbf{I}_{T\tilde{n}_T,T\tilde{n}_T}-\xi\mathbb{Z}(\mathbf{I}_{Tp,Tp}+\xi\mathbb{Z}^\top\mathbb{Z})^{-1}\mathbb{Z}^\top,$$

which, when using the othonormality assumption, simplifies to: $\mathbf{I}_{\tilde{n}_T,\tilde{n}_T}-\xi(1+\xi)^{-1}\mathbb{Z}\mathbb{Z}^\top$. Furthermore, note that $\mathbb{X}^\top\mathbb{Z}=\mathbb{I}_T^\top$ where $\mathbb{I}_T=\mathbf{1}_T\otimes\mathbf{I}_{pp}$ and $\mathbb{Z}^\top\mathbb{X}=\mathbb{I}_T$. Substitute this in the estimator to arrive at:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})} &= \{\mathbb{X}^\top[\mathbf{I}_{\tilde{n}_T,\tilde{n}_T}-\xi(1+\xi)^{-1}\mathbb{Z}\mathbb{Z}^\top]\mathbb{X}\}^{-1}\mathbb{X}^\top[\mathbf{I}_{\tilde{n}_T,\tilde{n}_T}-\xi(1+\xi)^{-1}\mathbb{Z}\mathbb{Z}^\top]\mathbb{Y} \\
&= [T\mathbf{I}_{pp}-\xi(1+\xi)^{-1}\mathbb{I}_T^\top\mathbb{I}_T]^{-1}[\mathbb{X}^\top\mathbb{Y}-\xi(1+\xi)^{-1}\mathbb{I}_T^\top\mathbb{Z}^\top\mathbb{Y}] \\
&= T^{-1}(1+\xi)[\mathbb{X}^\top\mathbb{Y}-\xi(1+\xi)^{-1}\mathbb{X}^\top\mathbb{Y}] \\
&= T^{-1}\mathbb{X}^\top\mathbb{Y} \;=\; T^{-1}(\mathbf{X}_1^\top\mathbf{Y}_1+\ldots+\mathbf{X}_T^\top\mathbf{Y}_T).
\end{aligned}$$

Clearly, $\mathbb{E}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})})$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})}) = T^{-1}(\sigma_\varepsilon^2+\sigma_\gamma^2)\mathbf{I}_{pp}$. Hence, $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})}) = [\mathbb{E}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})})-\boldsymbol{\beta}]^\top[\mathbb{E}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})})-\boldsymbol{\beta}] + \mathrm{tr}[\mathrm{Var}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})})] = pT^{-1}(\sigma_\varepsilon^2+\sigma_\gamma^2)$.

For the mean squared error of the updated ridge regression estimator, note that, when using the orthonormality assumption:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)] &= \mathbb{E}\{(\mathbf{X}_{t+1}^\top\mathbf{X}_{t+1}+\lambda_{t+1}\mathbf{I}_{pp})^{-1}[\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1}+\lambda_{t+1}\hat{\boldsymbol{\beta}}_t(\lambda_t)]\} \\
&= (1+\lambda_{t+1})^{-1}\{\mathbb{E}(\mathbf{X}_{t+1}^\top\mathbf{Y}_{t+1})+\lambda_{t+1}\mathbb{E}[\hat{\boldsymbol{\beta}}_t(\lambda_t)]\} \\
&= (1+\lambda_{t+1})^{-1}\{\boldsymbol{\beta}+\lambda_{t+1}\mathbb{E}[\hat{\boldsymbol{\beta}}_t(\lambda_t)]\}.
\end{aligned}$$

When the sequence of updated ridge regression estimators $\{\hat{\boldsymbol{\beta}}_{t+1}(\lambda_{t+1},\hat{\boldsymbol{\beta}}_t)\}_{t=1}^T$ is initiated with an unbiased estimator, all subsequent estimators are unbiased. Hence,

$$\mathrm{MSE}\{[\hat{\boldsymbol{\beta}}_T(\lambda_T)]\} \;=\; \mathrm{tr}\{\mathrm{Var}[\hat{\boldsymbol{\beta}}_T(\lambda_T)]\} \;=\; p\sigma_\varepsilon^2\sum_{t_h=1}^{T}\lambda_{t_h}^{-2}\prod_{\tau=t_h}^{T}\lambda_\tau^2(1+\lambda_\tau)^{-2},$$

where the expression of the variance is derived from that presented in the main body of the article with a general design matrix.

Now use the analytic expressions of the mean squared error of the two estimators and that $\lambda_t > T^{1/2}[\sigma_\varepsilon^{-2}(\sigma_\varepsilon^2 + \sigma_\gamma^2)]^{1/2}2^{t/2}$ for t such that $1 \leq t \leq T$:

$$\begin{aligned}
\mathrm{MSE}[\hat{\boldsymbol{\beta}}_T(\lambda_T)] &= p\sigma_\varepsilon^2 \sum_{t_h=1}^{T} \lambda_{t_h}^{-2} \prod_{\tau=t}^{T} \lambda_\tau^2 (1+\lambda_\tau)^{-2} \\
&< p\sigma_\varepsilon^2 \sum_{t=1}^{T} \lambda_t^{-2} &&< pT^{-1}(\sigma_\varepsilon^2 + \sigma_\gamma^2) \sum_{t=1}^{T} 2^{-T} \\
&< pT^{-1}(\sigma_\varepsilon^2 + \sigma_\gamma^2) &&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}_T^{(\mathrm{me})}),
\end{aligned}$$

as is claimed. □

SM III: Multi-target simulation

Here we investigate the use of the multiple targets. In particular, we assess whether the multi-targeted ridge regression estimator shrinks most towards the preferred target. Hereto we sample from the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the elements of the design matrix $\mathbf{X}$ drawn from the standard normal distribution, the elements of regression parameter as $\{\beta_j\}_{j=1}^{101} = \{(j - 51)/20\}_{j=1}^{101}$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{nn})$ with $n = 25$. We equip the estimator with the following three targets: $\boldsymbol{\beta}_0 = \frac{1}{4}$, $\boldsymbol{\beta}_0$ such that $\beta_{0,j} = -1$ if j is odd and $\beta_{0,j} = 1$ if j is even, and $\boldsymbol{\beta}_0 = 0_p$. Of these targets, the first is most informative. We employ 10-fold cross-validation to select the penalty parameters, one for each target. This is done 10.000 times, resulting in an equal number of triples $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)$. To assess to which target the estimator is shrunken most we plot the points $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)(\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3)^{-1} \in [0, 1]^3$ in a triangular plot. In addition, we provide a histogram of $\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3$ for an impression of the overall shrinkage. Both are provided in the top panels of Figure 1.

The whole exercise above has been repeated using a design matrix $\mathbf{X}$ with elements that are correlated and non-zero centered. In particular, the rows of $\mathbf{X}$ are sampled from the multivariate normal distribution $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is banded: $(\boldsymbol{\Sigma})_{jj} = 1$ for all j, $(\boldsymbol{\Sigma})_{j,j+1} = 0.5 = (\boldsymbol{\Sigma})_{j+1,j}$ for $j = 1, p-1$, $(\boldsymbol{\Sigma})_{j,j+2} = 0.25 = (\boldsymbol{\Sigma})_{j+2,j}$ for $j = 1, p-2$, $(\boldsymbol{\Sigma})_{j,j+3} = 0.1 = (\boldsymbol{\Sigma})_{j+3,j}$ for $j = 1, p-3$, and zero otherwise. Then, to all elements of each column an offset sampled from $\mathcal{U}[-10, 10]$ is added. All other aspects of the set-up are left unaltered. The resulting plots are shown in the lower panels of Figure 1.

The right panels of Figure 1 shows that virtuall all points fall close to the axis $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)(\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3)^{-1} = (1 - \alpha, \alpha, 0)$ with $\alpha \in [0, 1]$, and that the fast majority of these points concentrated either at or close to the point $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)(\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3)^{-1} = (1, 0, 0)$. This indicates that the largest weight are assigned to the most informative target, $\boldsymbol{\beta}_0 = \frac{1}{4}\boldsymbol{\beta}$. Hence, even in high-dimensional settings the method is able to assess the usefulness of a target from a set of targets for the problem at hand.

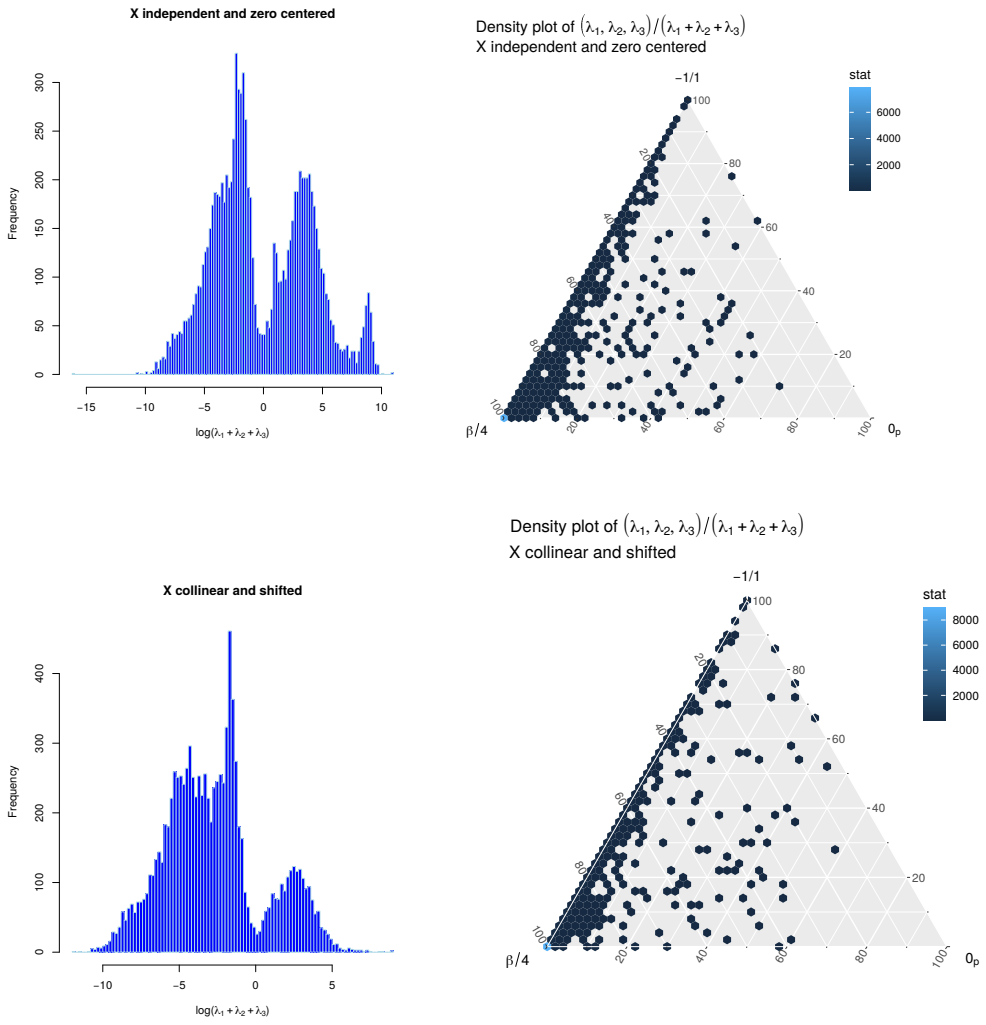

Figure 1: Computating times in $\log(\mathrm{ms})$ of the IRLS algorithm for various target choices, dimension p, and value of the penalty parameter. Computing times are obtained from 100 evaluations of the same data set using the `microbenchmark`-package [3]. This exercises has been repeated 100 times, each time with a different draw of the data. Each element of the random target is drawn from the standard normal distribution.

SM IV: IRLS algorithm computing times

We assess the effect of the target choice on the computation time of the IRLS algorithm for the evaluation of targeted ridge logistic regression estimator. Hereto we sample data from the logistic regression model with the elements of the design matrix all drawn from the standard normal distribution. Moreover, we set the elements of the regression parameter to $\{\beta_j\}_{j=1}^{101} = \{(j-51)/20\}_{j=1}^{101}$ or $\{\beta_j\}_{j=1}^{1001} = \{(j-501)/2000\}_{j=1}^{1001}$. The response of the i-th individual, Y_i with $i = 1, \ldots, 25$, is then drawn for the Bernoulli distribution with success probability $\exp(\mathbf{X}_{i,*}\boldsymbol{\beta})/[1+\exp(\mathbf{X}_{i,*}\boldsymbol{\beta})]$. With these data, we evaluate the targeted ridge logistic regression estimator for three choices of the penalty parameter $\lambda = \{1, 10, 100\}$ and six choices of the target: $\boldsymbol{\beta}_0 = \mathbf{0}_p$, $\boldsymbol{\beta}_0 = \frac{1}{2}\boldsymbol{\beta}$, $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$, $\boldsymbol{\beta}_0 = 2\boldsymbol{\beta}$, , $\boldsymbol{\beta}_0 = -\boldsymbol{\beta}$, and a random $\boldsymbol{\beta}_0$ with all its elements drawn from the standard normal distribution. The computation times of these evaluations are recorded by the `microbenchmark`-package [3]. The package does so a hundred times, shuffling the order of the evaluation the estimator with the different targets. As computing times may depend on the data, this whole exercise was repeated a hundred times. In total, for each $(p, \lambda, \boldsymbol{\beta}_0)$-combination a 10.000 computing times are available. This are displayed in Figure 2.

The boxplots in Figure 2 show that computing time of the IRLS algorithm for a zero-targeted ridge logistic regression estimator is overall smallest. For other targets, $\boldsymbol{\beta}_0 = \frac{1}{2}\boldsymbol{\beta}$, one or two extra iterations are needed for the IRLS algorithm to converge. Relatively, these extra iterations required for the evaluation with a zero target or $\boldsymbol{\beta}_0 = \frac{1}{2}\boldsymbol{\beta}$ do not lead to a dramatic increase in computing time.

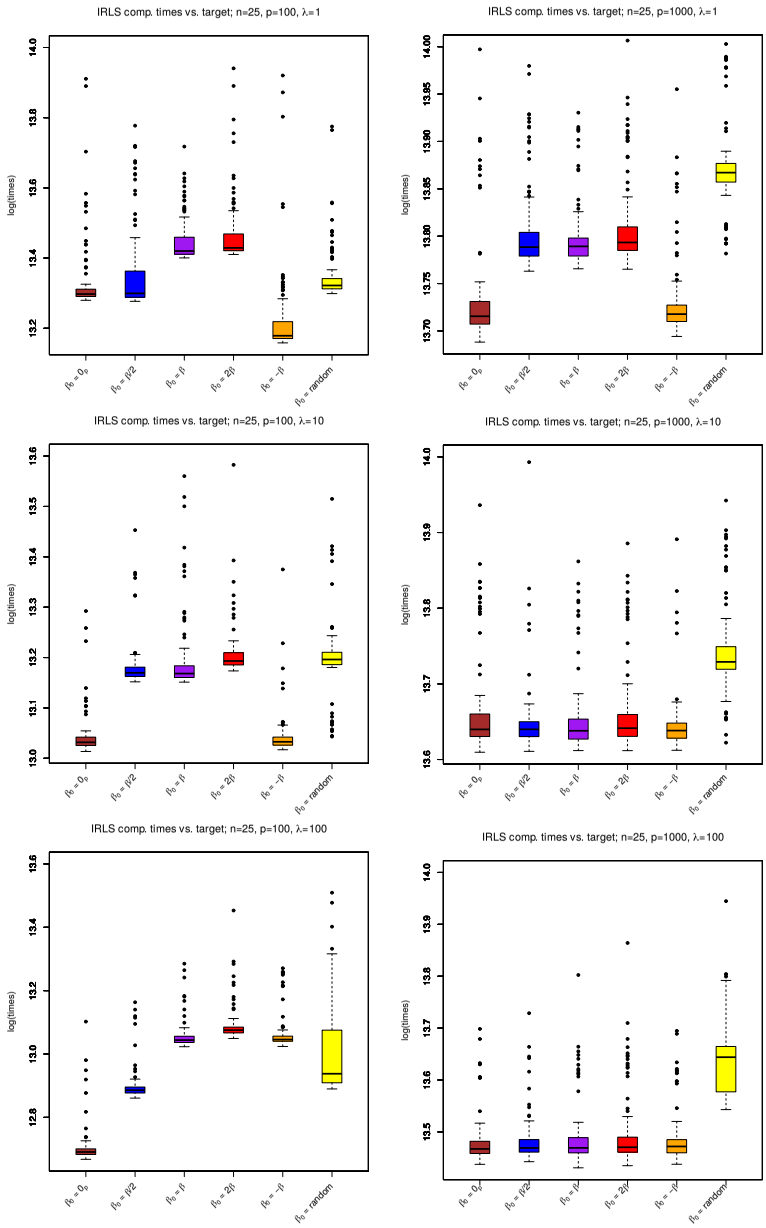

Figure 2: Computating times in log(ms) of the IRLS algorithm for various target choices, dimension p, and value of the penalty parameter. Computing times are obtained from 100 evaluations of the same data set using the `microbenchmark`-package [3]. This exercises has been repeated 100 times, each time with a different draw of the data. Each element of the random target is drawn from the standard normal distribution.

SM V: Constrained cross-validation

We compare constrained cross-validation to its unconstrained counterpart. Hereto 'historic' data are sampled from the linear regression model with parameter $\{\beta_j\}_{j=1}^{101} = \{(j-51)/20\}_{j=1}^{101}$, a design matrix with elements drawn from the standard normal distribution, and a unit error variance. The sample size of the historical data ranges from $n_{\text{past}} = 0$ to $n_{\text{past}} = 250$. The 'novel' or 'current' data, from which the parameter is estimated, are sampled with $n \in \{25, 50\}$ from either the same model or an empty one, i.e. with a zero regression parameter. Furthermore, the target is either informative or off, that is, $\boldsymbol{\beta}_0 = \frac{1}{2}\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0 = -\frac{1}{2}\boldsymbol{\beta}$, respectively. The penalty parameter is chosen by means of 10-fold cross-validation with various sample sizes of the historic data. The case $n_{\text{past}} = 0$ corresponds to the regular or unconstrained 10-fold cross-validation. With the cross-validated penalty parameter at hand, the loss of the estimate, $\|\hat{\boldsymbol{\beta}}(\lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}\|_2^2$, is evaluated. This exercise is repeated a hundred times, and the results are displayed in Figure 4.

The left panels of Figure 3 show that constrained cross-validation has a desirable effect (a reduction) on the loss if the current data come from a different model than the historic data. It safeguards against being swayed by the issues of the current data. If, however, the current data stem from the same model as the historic data, constrained cross-validation has no noticeable effect on the loss. The choice of the target, however, matters. The aforementioned desirable effect requires the target to be informative of the model from which the historic data stem. An off target renders no such effect of constrained cross-validation. In conclusion, constrained cross-validation in combination with an informative target safeguards against 'outlying' current data sets.

In an other comparison, the historical data are from the correct model but the novel data are from an empty model while the target is informative. The parameters and set-up are as above, except for $\boldsymbol{\beta} = \mathbf{0}_p$ in the generation of the novel data and either $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ or $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{\text{past}} = (\mathbf{X}_{\text{past}}^\top \mathbf{X}_{\text{past}})^+ \mathbf{X}_{\text{past}}^\top \mathbf{Y}_{\text{past}}$, where $\mathbf{A}^+$ denotes the Moore-Penrose generalized inverse. Again the penalty parameter is chosen by means of 10-fold constrained and unconstrained cross-validation with various samples sizes of the historical data. The ridge targeted regression estimator is evaluated with these cross-validated penalty parameters and the (logarithm of the) loss of the estimates calculated. This is repeated a hundred times, and the results are displayed in Figure 4.

The left panels of Figure 4, corresponding to a perfect target, shows that even with little historical data, constrained cross-validation is beneficial, but even more if the sample size of the historical data grows. For the estimated target, for small sample sizes of the historical data, there is little to no difference between the loss of the estimator with either an un- or constrained cross-validated penalty parameter. This is due to fact that even this target poor and does not perform well on historical data. For large sample sizes of the historical data (i.e $n_{\text{past}} > p$), the target becomes more informative for the historical data and the constraint kicks in, which result in a larger penalty parameter that shrinks more to the informative target and ignores the unrelated novel data. Then, constrained cross-validation is again more beneficial than its unconstrained counterpart. Note: the hick-up in the the right panel at $n_{\text{past}} \approx p$ is due to the ill-conditionedness of the matrix $\mathbf{X}_{\text{past}}^\top \mathbf{X}_{\text{past}}$.

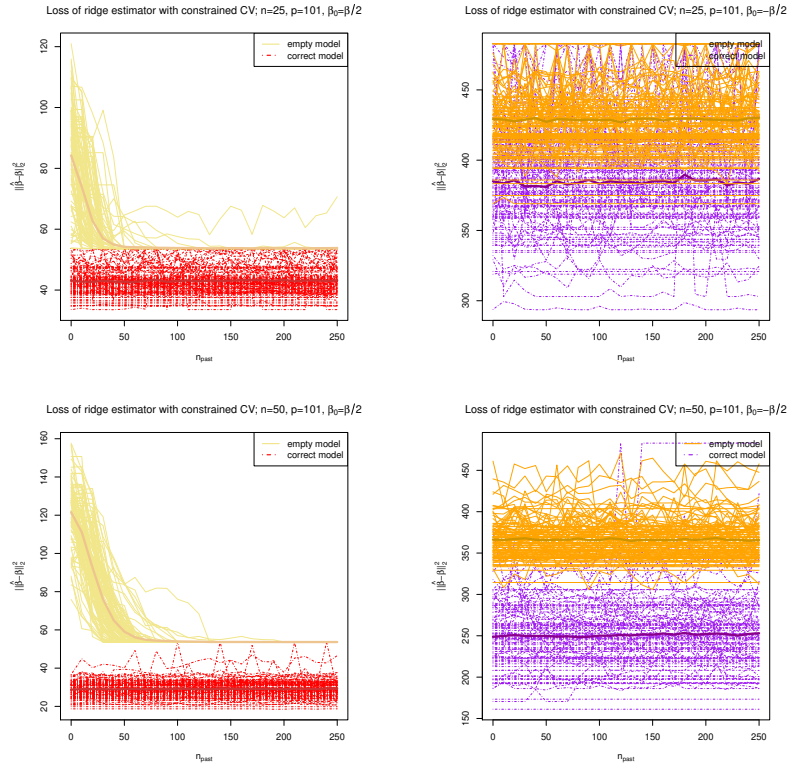

Figure 3: The loss of the ridge estimator with the penalty parameter chosen via constrained 10-fold cross-validation. The x-axis shows the sample size of the historical data, with $n_{\text{past}} = 0$ representing unconstrained cross-validation. The left and right panels show the results with an informative and off target, respectively. The top and bottom panel correspond to $n = 25$ and $n = 50$, respectively. Each panel shows the results with the current data sampled from the correct and empty model.

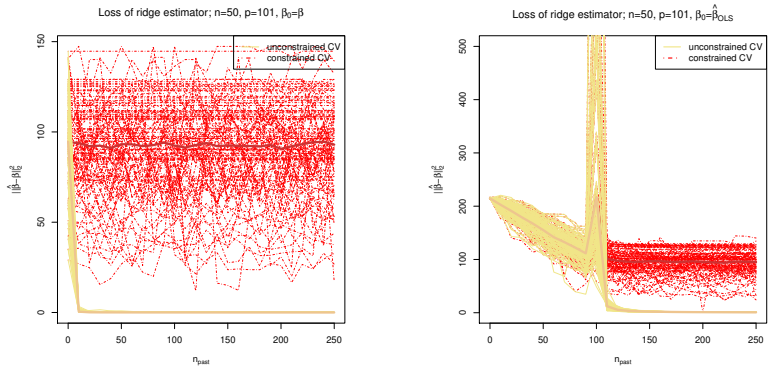

Figure 4: The loss of the ridge estimator with the penalty parameter chosen via constrained 10-fold un- and constrained cross-validation. The left and right panel correspond to the different informative targets $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ (left) and $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}$ (right). The x-axis shows the sample size of the historical data. Each panel shows the results with the current data sampled from the empty model, but the penalty parameter chosen differently.

SM VI: De novo simulation, additional plots

The '*de novo*' simulation of Section 5.1 of the main text is repeated. In the main text the elements of the design matrix are all sampled from the standard normal. Here three alternative ways generating the design matrix are employed:

i) The covariates are still independent, but are all differently centered. Hence, the location of a covariate differs over the sequence of data sets. Hereto the design matrix is generated as follows. First draw all elements of $\mathbf{X}$ from the standard normal. Then, add to all elements of each column an offset sampled from $\mathcal{U}[-10, 10]$.

ii) The covariates are zero centered, but no longer independent. Hereto the rows of $\mathbf{X}$ are sampled from the multivariate normal distribution $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is banded: $(\boldsymbol{\Sigma})_{jj} = 1$ for all j, $(\boldsymbol{\Sigma})_{j,j+1} = 0.5 = (\boldsymbol{\Sigma})_{j+1,j}$ for $j = 1, p-1$, $(\boldsymbol{\Sigma})_{j,j+2} = 0.25 = (\boldsymbol{\Sigma})_{j+2,j}$ for $j = 1, p-2$, $(\boldsymbol{\Sigma})_{j,j+3} = 0.1 = (\boldsymbol{\Sigma})_{j+3,j}$ for $j = 1, p-3$, and zero otherwise.

iii) The covariates are nonzero centered and dependent. This is a combination of the previous two case. First, the row of $\mathbf{X}$ are sampled from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ as above. Then, an offset is added to each covariate, as above.

All other aspects of the simulation set-up, e.g. the sample size, dimension, choice of the parameters are unchanged. The results are shown in Figure 5. Figure 5 contains, for reference, also the results of the original simulation.

Figure 5 shows that a change in the location of the covariates does not affect the updating, why dependency among the covariates does. In particular, dependency slows down the convergence to the true parameter value.

We have also repeated the '*de novo*' simulation of Section 5.1 of the main text but with different initial targets, now using $\boldsymbol{\beta}_0 = -\boldsymbol{\beta}$, $\boldsymbol{\beta}_0 = \frac{1}{2}\boldsymbol{\beta}$, $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$, and , $\boldsymbol{\beta}_0 = 2\boldsymbol{\beta}$. The results, shown in Figure 6, are in line with intuition: the closer the initial target to the true parameter, the better the performance.

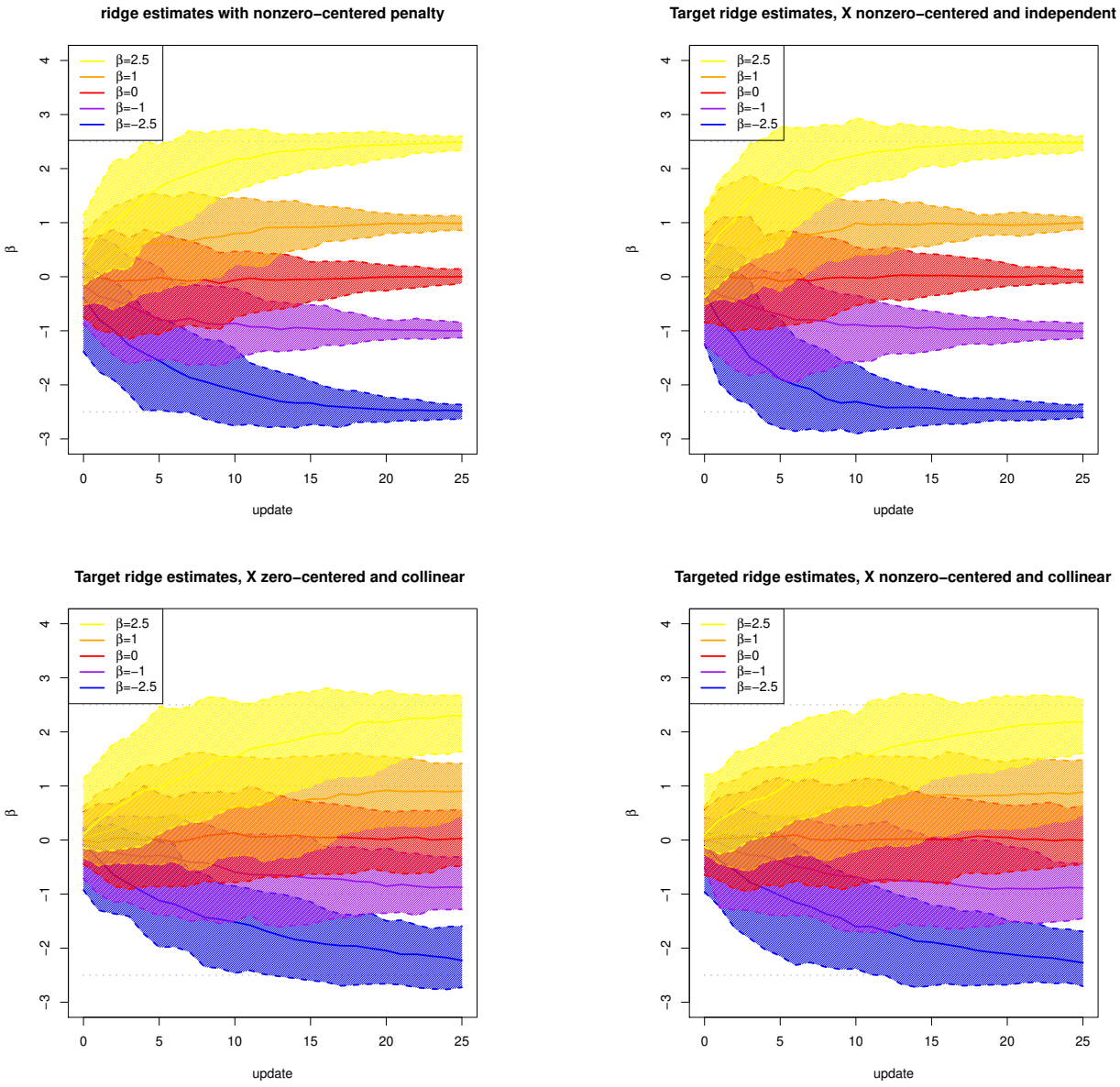

Figure 5: The panels show the $(5\%, 95\%$-quantile intervals of the updated ridge estimates of β_j with $j \in \{0, 30, 50, 70, 100\}$ plotted against t for regression models with different design matrices. The solid, colored line inside these intervals is the corresponding 50% quantile. The dotted, grey lines are the true values of the β_j's.

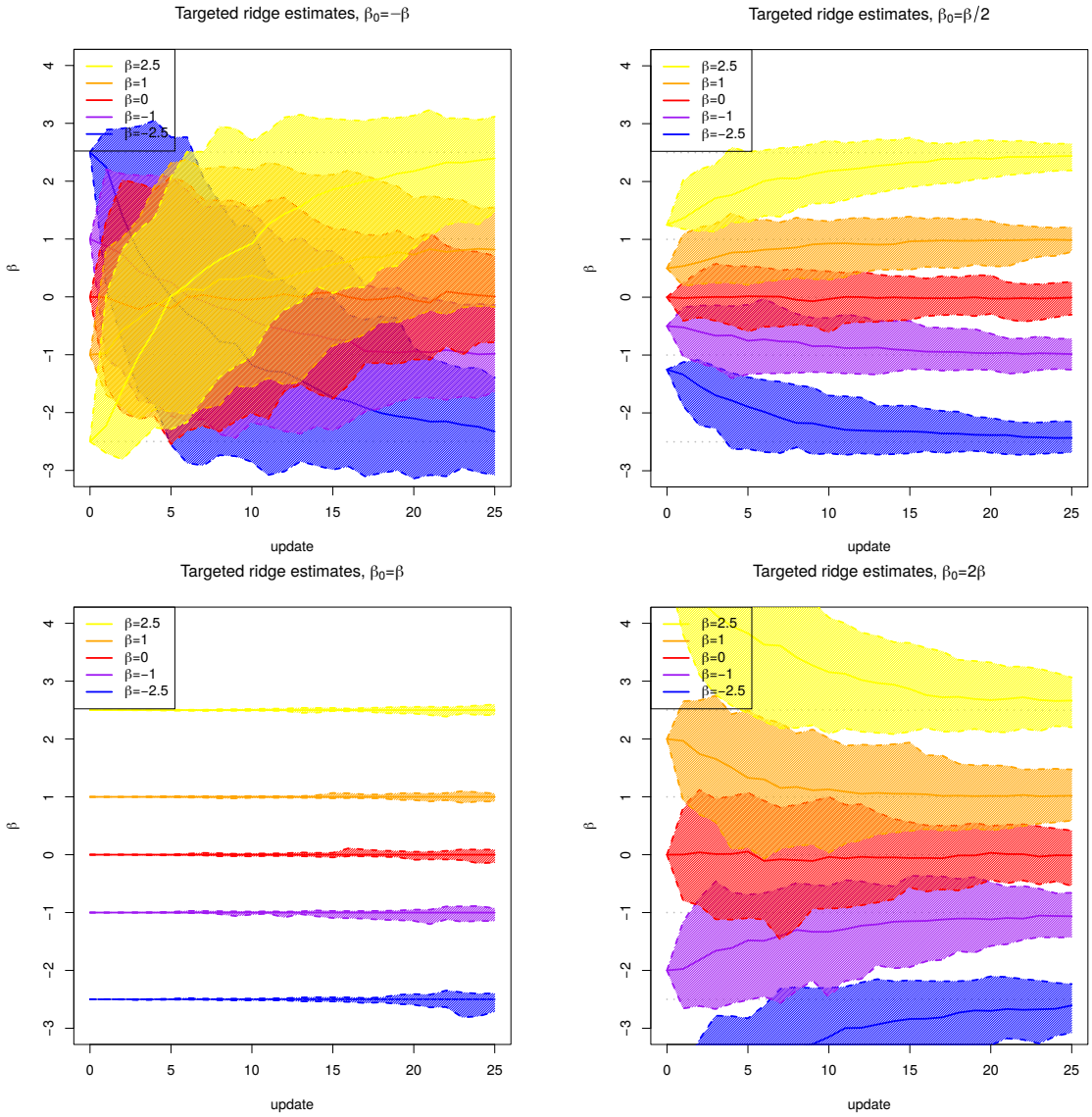

Figure 6: The panels show the $(5\%, 95\%$-quantile intervals of the updated ridge estimates of β_j with $j \in \{0, 30, 50, 70, 100\}$ plotted against t for regression models with different initial targets. The solid, colored line inside these intervals is the corresponding 50% quantile. The dotted, grey lines are the true values of the β_j's.

SM VII: Application

year	# counties	# health indicators
2008	159	1
2009	159	1
2010	158	8
2011	111	16
2012	96	17
2013	79	22
2014	60	23
2015	67	22
2016	57	23

Table 1: Table of number of counties with full information on the response (suicide rate) and the registered health indicators, with their number in the last column, per year.

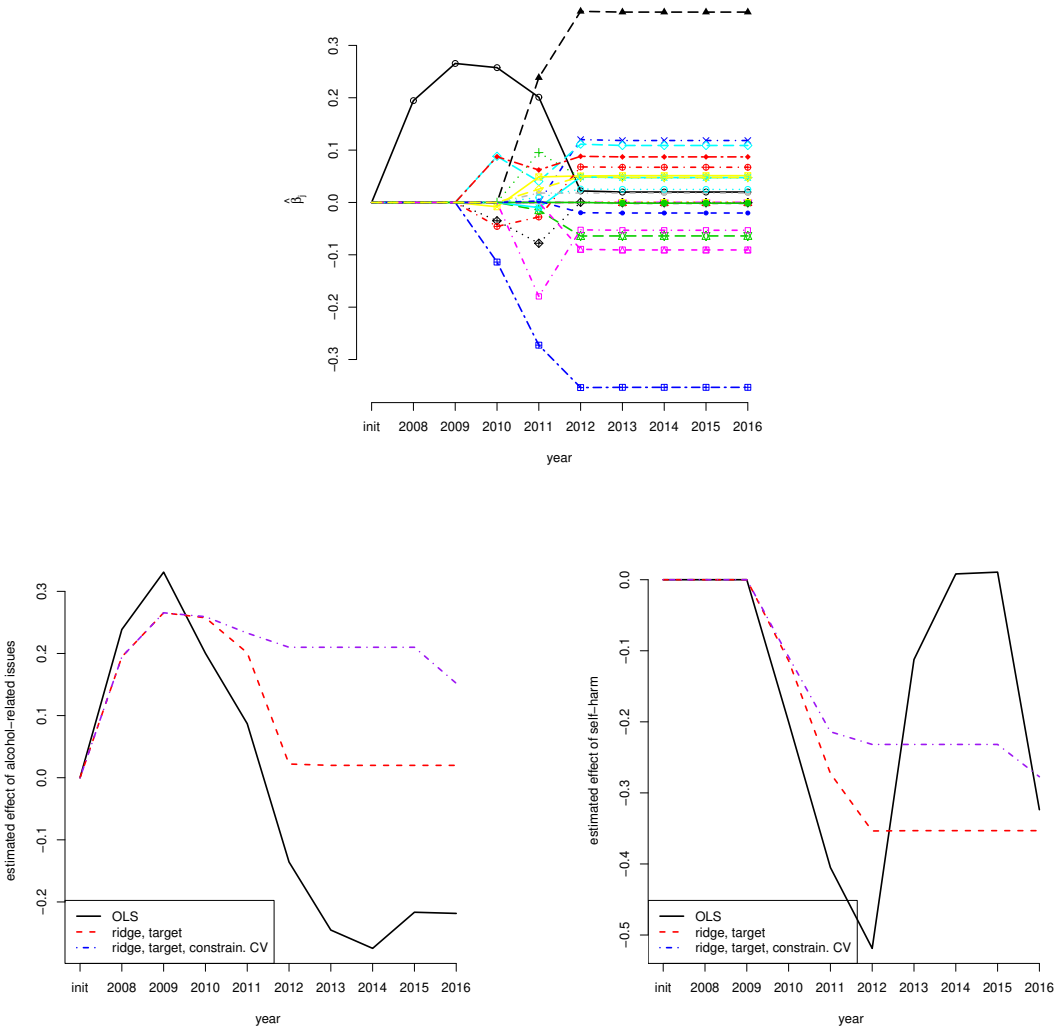

Figure 7: The top panel shows the trajectories of the the updated ridge regression, with its penalty parameter chosen via unconstrained LOOCV, estimates. Each trajectory represents a single covariate. The presence of a health indicator in the data of a particular year is evident from a symbol on its trajectory at the corresponding year. The symbol is omitted in years that the health indicator was not registered. The differences among the different estimators' trajectories of the same regression coefficient is highlighted for the two covariates with the largest effects (lower panels).

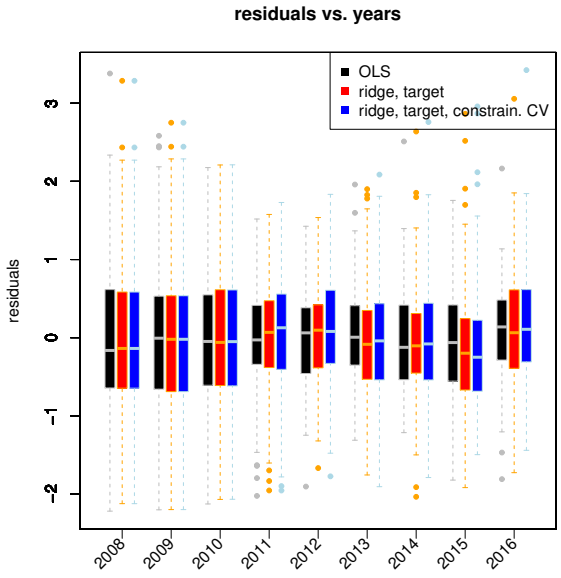

Figure 8: Residuals of the three fits, maximum likelihood (OLS) and the updated ridge regression, with its penalty parameter chosen via un- and constrained LOOCV, vs. year.

References

[1] D A Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, New York, 2008.

[2] E L Lehmann and G Casella. *Theory of Point Estimation*. Springer, 2006.

[3] O Mersmann. *microbenchmark: Accurate Timing Functions*, 2014. R package version 1.4-2.

[4] J Stachurski. *Economic Dynamics: Theory and Computation*. MIT Press, 2009.

[5] A W van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.