Journal of Machine Learning Research 21 (2020) 1-24 Submitted 10/15; Revised 8/19; Published 3/20

# Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes — Supplementary Material —

**Anders Ellern Bilgrau**★ ANDERS.ELLERN.BILGRAU@GMAIL.COM
*Department of Mathematical Sciences,*
*Aalborg University*
*9220 Aalborg Ø, Denmark &*
*Department of Haematology,*
*Aalborg University Hospital*
*9000 Aalborg, Denmark*

**Carel F.W. Peeters**★ CF.PEETERS@AMSTERDAMUMC.NL
*Department of Epidemiology & Biostatistics,*
*Amsterdam University medical centers, location VUmc*
*Postbus 7057, 1007 MB Amsterdam, The Netherlands*

**Poul Svante Eriksen** SVANTE@MATH.AAU.DK
*Department of Mathematical Sciences,*
*Aalborg University*
*9220 Aalborg Ø, Denmark*

**Martin Bøgsted** M_BOEGSTED@DCM.AAU.DK
*Department of Haematology,*
*Aalborg University Hospital*
*9000 Aalborg, Denmark &*
*Department of Clinical Medicine,*
*Aalborg University*
*9000 Aalborg, Denmark*

**Wessel N. van Wieringen** W.VANWIERINGEN@AMSTERDAMUMC.NL
*Department of Epidemiology & Biostatistics,*
*Amsterdam University medical centers, location VUmc*
*Postbus 7057, 1007 MB Amsterdam, The Netherlands &*
*Department of Mathematics,*
*VU University Amsterdam*
*1081 HV Amsterdam, The Netherlands*

**Editor:** Francis Bach

This supplement is structured as follows: Section 1 gives alternative updating schemes for obtaining the precision estimates. Section 2 gives details on estimation in certain special cases. Section 3 derives an approximation to the fused leave-one-out cross-validation score. Section 4 gives the remainder of the results for Simulation Scenario 2. Section 5 gives the

★. Shared first authorship.

remainder of the results for Simulation Scenario 5. Last, Section 6 gives the remainder of the results for Simulation Scenario 6.

## 1. Alternative Fused Ridge Solutions

This section derives two equivalent (in terms of Equation 7) alternative updating schemes to (8). The motivation for the exploration of these alternative recursive estimators is twofold. First, alternative recursions can exhibit differing numerical (in)stability for extreme values of the penalty matrix $\mathbf{\Lambda} = [\lambda_{g_1 g_2}]$. Second, they provide additional intuition and understanding of the targeted fused ridge estimator.

The general strategy to finding the alternatives is to rewrite the gradient equation (27) into the non-fused form (28), which we will repeat here:

$$\hat{\mathbf{\Omega}}_{g_0}^{-1} - \bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}(\hat{\mathbf{\Omega}}_{g_0} - \bar{\mathbf{T}}_{g_0}) = \mathbf{0}, \tag{S1}$$

where $\bar{\lambda}_{g_0}$, $\bar{\mathbf{T}}_{g_0}$, and $\bar{\mathbf{S}}_{g_0}$ do not depend on $\hat{\mathbf{\Omega}}_{g_0}$. Note that an explicit closed-form solution to (S1) exists in the form of (7).

### 1.1 First Alternative

The first alternative scheme is straightforward. Rewrite (27) to:

$$\begin{aligned}\mathbf{0} &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}(\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g\neq g_0}\lambda_{gg_0}(\mathbf{\Omega}_g - \mathbf{T}_g) \\ &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left\{\hat{\mathbf{\Omega}}_{g_0} - \left[\mathbf{T}_{g_0} + \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{\lambda_{g_0\bullet}}(\mathbf{\Omega}_g - \mathbf{T}_g)\right]\right\},\end{aligned} \tag{S2}$$

where $\lambda_{g_0\bullet} = \sum_g \lambda_{gg_0}$. In terms of (S1), we thus have the updating scheme given in equation (9). As stated in the main text, it has the intuitive interpretation that a fused class target is used which is a combination of the class-specific target and the 'target corrected' estimates of remaining classes.

### 1.2 Second Alternative

We now derive a second alternative recursion scheme. Add and subtract $\lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g$, to (S2) and rewrite such that:

$$\begin{aligned}\mathbf{0} &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}(\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g + \sum_{g\neq g_0}\lambda_{gg_0}(\mathbf{\Omega}_g - \mathbf{T}_g) - \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g \\ &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g\right)\right] + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g - \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{T}_g - \lambda_{g_0\bullet}\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g \\ &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\mathbf{S}_{g_0} - \lambda_{g_0\bullet}\left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g\right)\right] - \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{T}_g - (\lambda_{g_0\bullet}-1)\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g \\ &= n_{g_0}\hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0}\left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}}\sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g + \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g\right] - \lambda_{g_0\bullet}\left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g\right)\right].\end{aligned}$$

Dividing by $n_{g_0}$ gives

$$\mathbf{0} = \hat{\mathbf{\Omega}}_{g_0}^{-1} - \left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}} \sum_{g\neq g_0} \lambda_{gg_0}\mathbf{\Omega}_g + \sum_{g\neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g\right] - \frac{\lambda_{g_0\bullet}}{n_{g_0}}\left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g\right)\right],$$

which brings the expression to the desired form (S1) with the updating scheme

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} + \frac{\lambda_{g_0\bullet}-1}{n_{g_0}} \sum_{g\neq g_0} \lambda_{gg_0}\mathbf{\Omega}_g + \sum_{g\neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}}\mathbf{T}_g, \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0} + \sum_{g\neq g_0}\lambda_{gg_0}\mathbf{\Omega}_g, \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0\bullet}}{n_{g_0}}.$$

Again, a solution for $\hat{\mathbf{\Omega}}_{g_0}$ with fixed $\mathbf{\Omega}_g$ for all $g \neq g_0$, is available through Lemma 8 (van Wieringen and Peeters, 2016) and is given in (7).

### 1.3 Motivation

Though seemingly more complicated, these alternative updating schemes can be numerically more stable for extreme penalties. In both alternatives, we see that $\bar{\mathbf{S}}_{g_0}$ is positive semi-definite for (nearly) all very large and very small penalties. Likewise, $\bar{\mathbf{T}}_{g_0}$ is always positive definite. Compare the alternative expressions to the updating scheme given by (8) which can be seen to be numerically unstable for very large penalties: For very large $\lambda_{gg}$ or $\lambda_{g_1g_2}$ the $\bar{\mathbf{S}}_{g_0}$ in (8) may be a matrix with numerically extreme values. This implies ill-conditioning and numerical instability under finite computer precision. On the other hand, 'updating' the target matrix will generally lead to updates for which the resulting estimator is not rotationally equivariant. This implies a reduction in computational speed.

## 2. Estimation in Special Cases

Here we explore scenarios for which we arrive at explicit targeted fused ridge estimators. These explicit solutions further insight into the behavior of the general estimator and they can provide computational speed-ups in certain situations. Three special cases are covered:

I. $\lambda_{gg'} = 0$ for all $g \neq g'$ or equivalently $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all $g$;

II. $\mathbf{\Omega}_1 = \cdots = \mathbf{\Omega}_G$ and $\mathbf{T}_g = \mathbf{T}$ for all $g$;

III. $\mathbf{T}_g = \mathbf{T}$ for all $g$, $\lambda_{gg} = \lambda$ for all $g$, $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$, and $\lambda_f \to \infty^-$.

### 2.1 Special Case I

When $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all $g$, we have that $\sum_{g'\neq g} \lambda_{gg'} = \sum_{g'\neq g} \lambda_{g'g} = 0$ for all $g$. Hence, all fusion penalties are zero. The zero gradient equation (27) for class $g$ then no longer hinges upon information from the remaining classes $g'$. The targeted fused precision estimate for class $g$ then reduces to (29) of Corollary 11. This case thus coincides, as expected, with obtaining $G$ decoupled non-fused ridge precision estimates. A special case that results in the same estimates occurs when considering $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$ and $\lambda_f$ is taken to be 0.

### 2.2 Special Case II

Suppose $\mathbf{\Omega}_g = \mathbf{\Omega}$ and $\mathbf{T}_g = \mathbf{T}$ for all $g$. Consequently, the fusion penalty term vanishes irrespective of the values of the $\lambda_{g_1 g_2}$, $g_1 \neq g_2$. The zero gradient equation (27) then reduces to

$$\mathbf{0} = n_g \hat{\mathbf{\Omega}}^{-1} - n_g \mathbf{S}_g - \lambda_{gg}(\hat{\mathbf{\Omega}} - \mathbf{T}),$$

for each class $g$. Adding all $G$ equations implies:

$$\begin{aligned}
\mathbf{0} &= \sum_{g=1}^{G} n_g \hat{\mathbf{\Omega}}^{-1} - \sum_{g=1}^{G} n_g \mathbf{S}_g - \left( \sum_{g=1}^{G} \lambda_{gg} \right) (\hat{\mathbf{\Omega}} - \mathbf{T}) \\
&= n_\bullet \hat{\mathbf{\Omega}}^{-1} - n_\bullet \mathbf{S}_\bullet - \operatorname{tr}(\mathbf{\Lambda})(\hat{\mathbf{\Omega}} - \mathbf{T}) \\
&= \hat{\mathbf{\Omega}}^{-1} - \left[ \mathbf{S}_\bullet - \frac{\operatorname{tr}(\mathbf{\Lambda})}{n_\bullet} \mathbf{T} \right] - \frac{\operatorname{tr}(\mathbf{\Lambda})}{n_\bullet} \hat{\mathbf{\Omega}}.
\end{aligned} \tag{S3}$$

We recognize that (S3) is of the form (22). Lemma 8 may then be directly applied to obtain the solution:

$$\hat{\mathbf{\Omega}}(\mathbf{\Lambda}) = \left\{ \left[ \lambda^* \mathbf{I}_p + \frac{1}{4}(\mathbf{S}_\bullet - \lambda^* \mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S}_\bullet - \lambda^* \mathbf{T}) \right\}^{-1}, \tag{S4}$$

where $\lambda^* = \operatorname{tr}(\mathbf{\Lambda})/n_\bullet$. Hence, this second special case gives a non-fused penalized estimate that uses the pooled covariance matrix. It can be interpreted as an averaged penalized estimator. It is of importance in testing equality of the class precision matrices (see Section 4.1 of the main text).

### 2.3 Special Case III

Suppose that $\mathbf{T}_g = \mathbf{T}$ for all $g$, that $\lambda_{gg} = \lambda$ for all $g$, and that $\lambda_{g_1 g_2} = \lambda_f$ for all $g_1 \neq g_2$. The main optimization problem then reduces to (6). Clearly, for $\lambda_f \to \infty^-$ the fused penalty

$$f^{\mathrm{FR}}(\{\mathbf{\Omega}_g\}; \lambda, \lambda_f, \mathbf{T}) = \frac{\lambda}{2} \sum_g \left\| \mathbf{\Omega}_g - \mathbf{T} \right\|_F^2 + \frac{\lambda_f}{4} \sum_{g_1, g_2} \left\| (\mathbf{\Omega}_{g_1} - \mathbf{\Omega}_{g_2}) \right\|_F^2$$

is minimized when $\mathbf{\Omega}_1 = \mathbf{\Omega}_2 = \cdots = \mathbf{\Omega}_G$. This is also implied, more rigorously, by Corollary 13. Hence, the problem reduces to the special case of section 2.2 considered above. The solution to the penalized ML problem when $\lambda_f = \infty$ is then given by (S4) where $\operatorname{tr}(\mathbf{\Lambda})$ now implies $G\lambda$.

## 3. Fused Kullback-Leibler Approximate Cross-Validation

### 3.1 Motivation

In $\ell_1$-penalized estimation of the precision matrix, penalty selection implies (graphical) model selection: Regularization results in automatic selection of conditional dependencies. One then seeks to select an optimal value for the penalty parameter in terms of model selection consistency. To this end, the Bayesian information criterion (BIC), the extended

BIC (EBIC), and the stability approach to regularization selection (StARS) are appropriate (Liu et al., 2010). The (fused) $\ell_2$-penalty will not directly induce sparsity in precision matrix estimates. Hence, in $\ell_2$-penalized problems it is natural to choose the penalty parameters on the basis of efficiency loss. Of interest are then estimators of the Kullback-Leibler (KL) divergence, such as LOOCV, generalized approximate cross-validation (GACV), and Akaike's information criterion (AIC). While superior in terms of predictive accuracy due to its data-driven nature, the LOOCV is computationally very expensive. Vujačić et al. (2015) proposed a KL-based CV loss with superior performance to both AIC and GACV. The proposed method has closed-form solutions and thus provides a fast approximation to LOOCV. Here, we extend this method to provide a computationally friendly approximation of the fused LOOCV score.

### 3.2 Formulation

Following Vujačić et al. (2015), we now restate the KL approximation to LOOCV in the fused ridge setting. Let the true precision matrix for class $g$ be denoted by $\boldsymbol{\Omega}_g$. Its estimate, shorthanded by $\hat{\boldsymbol{\Omega}}_g$ can be obtained through Algorithm 1. The KL divergence between the multivariate normal distributions $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}_g^{-1})$ and $\mathcal{N}_p(\mathbf{0}, \hat{\boldsymbol{\Omega}}_g^{-1})$ can be shown to be:

$$\mathrm{KL}(\boldsymbol{\Omega}_g, \hat{\boldsymbol{\Omega}}_g) = \frac{1}{2}\Big\{\mathrm{tr}(\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g) - \ln|\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g| - p\Big\}.$$

For each $g$ we wish to minimize this divergence. In the fused case we therefore consider the *fused Kullback-Leibler* (FKL) divergence which, motivated by the LOOCV score, is taken to be a weighted average of KL divergences:

$$\begin{aligned}&\mathrm{FKL}(\{\boldsymbol{\Omega}_g\}, \{\hat{\boldsymbol{\Omega}}_g\}) \\ &\quad= \frac{1}{n_\bullet}\sum_{g=1}^{G} n_g \,\mathrm{KL}(\boldsymbol{\Omega}_g, \hat{\boldsymbol{\Omega}}_g) = \frac{1}{n_\bullet}\sum_{g=1}^{G} \frac{n_g}{2}\Big\{\mathrm{tr}(\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g) - \ln|\boldsymbol{\Omega}_g^{-1}\hat{\boldsymbol{\Omega}}_g| - p\Big\}. \qquad \text{(S5)}\end{aligned}$$

The FKL divergence (S5) can, using the likelihood (3), be rewritten as

$$\mathrm{FKL} = -\frac{1}{n_\bullet}\mathcal{L}\big(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}\big) + \mathrm{bias}, \quad \text{where} \quad \mathrm{bias} = \frac{1}{2n_\bullet}\sum_{g=1}^{G} n_g \,\mathrm{tr}\big[\hat{\boldsymbol{\Omega}}_g(\boldsymbol{\Omega}_g^{-1} - \mathbf{S}_g)\big],$$

and where the equality holds up to the addition of a constant. It is clear that the bias term depends on the unknown true precision matrices and thus needs to be estimated. The fused analogue to the proposal of Vujačić et al. (2015), called the *fused Kullback-Leibler approximate cross-validation* score or simply *approximate fused LOOCV* score, then is

$$\widehat{\mathrm{FKL}}(\boldsymbol{\Lambda}) = -\frac{1}{n_\bullet}\mathcal{L}\big(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}\big) + \widehat{\mathrm{bias}}, \qquad \text{(S6)}$$

with

$$\widehat{\mathrm{bias}} = \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big\{\mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}_g^2 - \hat{\boldsymbol{\Omega}}_g)\mathbf{y}_{ig} + \bar{\lambda}_g\mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}_g^4 - \hat{\boldsymbol{\Omega}}_g^3)\mathbf{y}_{ig}\Big\}, \qquad \text{(S7)}$$

and where $\bar{\lambda}_g = \frac{\lambda_{g\bullet}}{n_g}$. The derivation of this estimate is given in Section 3.3 below. One would then choose $\boldsymbol{\Lambda}^\star$ such that the FKL approximate cross-validation score is minimized:

$$\boldsymbol{\Lambda}^\star = \arg\min_{\boldsymbol{\Lambda}} \widehat{\mathrm{FKL}}(\boldsymbol{\Lambda}), \quad \text{subject to:} \quad \boldsymbol{\Lambda} \geq \mathbf{0} \wedge \mathrm{diag}(\boldsymbol{\Lambda}) > \mathbf{0}. \tag{S8}$$

The closed form expression in (S6) implies that $\boldsymbol{\Lambda}^\star$ is more rapidly determined than $\boldsymbol{\Lambda}^*$. As seen in the derivation, $\boldsymbol{\Lambda}^* \approx \boldsymbol{\Lambda}^\star$ for large sample sizes.

### 3.3 Derivation

Here we give, borrowing some ideas from Vujačić et al. (2015), the derivation of the estimate (S6). Let observation $i$ in class $g$ be denoted by $\mathbf{y}_{ig}$ and let $\mathbf{S} = \mathbf{S}_{ig} = \mathbf{y}_{ig}\mathbf{y}_{ig}^\top$ be the sample covariance or scatter matrix of that observation. As before, the singularly indexed $\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} \mathbf{S}_{ig}$ is the class-specific sample covariance matrix. Throughout this section we will conveniently drop (some of) the explicit notation.

The FKL divergence reframes the LOOCV score in terms of a likelihood evaluation and a bias term when $\mathbf{S}$ is *not* left out of class $g$. We thus study the change in the estimate as function of the single scatter matrix $\mathbf{S}$. Let $\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) = \hat{\boldsymbol{\Omega}}_g^{\neg ig}$ be the estimate in class $g$ when $\mathbf{S}$ is omitted. That is, $\hat{\boldsymbol{\Omega}}_g(\mathbf{S})$ is part of the solution to the system

$$\boldsymbol{\Omega}_a^{-1} + \mu_{aa}\boldsymbol{\Omega}_a + \mathbb{1}[a{=}g]\mathbf{S} + \sum_{b\neq a} \mu_{ab}\boldsymbol{\Omega}_b + \mathbf{A}_a = \mathbf{0}, \quad \text{for all} \quad a = 1, \ldots, G, \tag{S9}$$

where $\mu_{aa} = -\frac{\lambda_{a\bullet}}{n_a}$, $\mu_{ab} = \frac{\lambda_{ab}}{n_a}$, and where $\mathbf{A}_a$ is a matrix determined by the remaining data, penalty parameters and targets. Note that the penalized MLE can be denoted $\hat{\boldsymbol{\Omega}}_g = \hat{\boldsymbol{\Omega}}_g(\mathbf{0})$, which corresponds to the 'full' estimate resulting from the full gradient equation (27).

We wish to approximate $\hat{\boldsymbol{\Omega}}_g(\mathbf{S})$ by a Taylor expansion around $\hat{\boldsymbol{\Omega}}_g(\mathbf{0})$, i.e.:

$$\hat{\boldsymbol{\Omega}}_a(\mathbf{S}) \approx \hat{\boldsymbol{\Omega}}_a(\mathbf{0}) + \sum_{j,j'} \frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}} S_{jj'}.$$

Differentiating (S9) w.r.t. $S_{jj'}$, the $(j, j')$th entry in $\mathbf{S}$, and equating to zero yields

$$\begin{aligned}
\mathbf{0} &= -\hat{\boldsymbol{\Omega}}_a^{-1}\frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}}\hat{\boldsymbol{\Omega}}_a^{-1} + \mu_{aa}\frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}} + \mathbb{1}[a{=}g]\mathbf{E}_{jj'} + \sum_{b\neq a}\mu_{ab}\frac{\partial \hat{\boldsymbol{\Omega}}_b}{\partial S_{jj'}} \\
&= -\hat{\boldsymbol{\Omega}}_a^{-1}\frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}}\hat{\boldsymbol{\Omega}}_a^{-1} + \sum_{b}\mu_{ab}\frac{\partial \hat{\boldsymbol{\Omega}}_b}{\partial S_{jj'}} + \mathbb{1}[a{=}g]\mathbf{E}_{jj'}, \quad \text{for all} \quad j, j',
\end{aligned} \tag{S10}$$

where $\mathbf{E}_{jj'}$ is the null matrix except for unity in entries $(j, j')$ and $(j', j)$. The third term is obtained as $\partial\mathbf{S}/\partial S_{jj'} = \mathbf{E}_{jj'}$ by the symmetric structure of $\mathbf{S}$. This is also seen from the fact that $\mathbf{S} = \sum_{jj'} S_{jj'}\mathbf{E}_{jj'}$. Let

$$\mathbf{V}(\mathbf{S})_a = \sum_{j,j'} \frac{\partial \hat{\boldsymbol{\Omega}}_a}{\partial S_{jj'}} S_{jj'},$$

and multiply (S10) by $S_{jj'}$ and sum over all $j, j'$ to obtain

$$\hat{\mathbf{\Omega}}_a^{-1}\mathbf{V}(\mathbf{S})_a\hat{\mathbf{\Omega}}_a^{-1} - \sum_b \mu_{ab}\mathbf{V}(\mathbf{S})_b = \mathbb{1}[a{=}g]\mathbf{S}, \quad \text{for all} \quad a = 1, \ldots, G. \tag{S11}$$

We seek the solution vector $\mathbf{V} = \big\{\mathbf{V}(\mathbf{S})_a\big\}_{a=1}^G$ of square matrices for the system of equations in (S11) which can be rewritten in the following way. Introduce and consider the linear operator (or block matrix):

$$\mathbf{N} = \big\{\mathbf{N}_{ab}\big\}_{a,b=1}^G \quad \text{where} \quad \mathbf{N}_{ab} = \begin{cases} \hat{\mathbf{\Omega}}_a^{-1} \otimes \hat{\mathbf{\Omega}}_a^{-1} - \mu_{aa}\mathbf{I}_p \otimes \mathbf{I}_p & \text{if} \quad a = b \\ -\mu_{ab}\mathbf{I}_p \otimes \mathbf{I}_p & \text{if} \quad a \neq b \end{cases}.$$

Then $\mathbf{V}$ can be verified to be the solution to the system (S10) as

$$\begin{aligned} \mathbf{N}(\mathbf{V})_a &= \sum_b \mathbf{N}_{ab}\mathbf{V}(\mathbf{S})_b = \mathbf{0} \quad \text{for} \quad a \neq g, \quad \text{and} \\ \mathbf{N}(\mathbf{V})_g &= \sum_b \mathbf{N}_{gb}\mathbf{V}(\mathbf{S})_b = \mathbf{S} \quad \text{for} \quad a = g. \end{aligned}$$

Hence we need to invert $\mathbf{N}$ to solve for $\mathbf{V}$. The structure of $\mathbf{N}$ is relatively simple, but there seems to be no (if any) simple inverse. Note that $\mathbf{N} = \mathbf{D} - \mathbf{M}$ is the difference of a (block) diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{M}$ depending on the $\mu$'s:

$$\begin{aligned} \mathbf{D}_{aa} &= \hat{\mathbf{\Omega}}_a^{-1} \otimes \hat{\mathbf{\Omega}}_a^{-1}, \\ \mathbf{M}_{ab} &= \mu_{ab}\mathbf{I}_p \otimes \mathbf{I}_p. \end{aligned}$$

In terms of the $\mu$'s we obtain to first order that

$$\mathbf{N}^{-1} = (\mathbf{D} - \mathbf{M})^{-1} \approx \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1},$$

yielding the approximation

$$\begin{aligned} \hat{\mathbf{\Omega}}_g(\mathbf{S}) &\approx \hat{\mathbf{\Omega}}_g + (\hat{\mathbf{\Omega}}_g \otimes \hat{\mathbf{\Omega}}_g + \mu_{gg}\hat{\mathbf{\Omega}}_g^2 \otimes \hat{\mathbf{\Omega}}_g^2)(\mathbf{S}) \\ &= \hat{\mathbf{\Omega}}_g + \hat{\mathbf{\Omega}}_g\mathbf{S}\hat{\mathbf{\Omega}}_g + \mu_{gg}\hat{\mathbf{\Omega}}_g^2\mathbf{S}\hat{\mathbf{\Omega}}_g^2, \end{aligned} \tag{S12}$$

where $\hat{\mathbf{\Omega}}_g = \hat{\mathbf{\Omega}}(\mathbf{0})$. To a first order in $\mu_{gg}$ this is the same as the approximation

$$\hat{\mathbf{\Omega}}_g(\mathbf{S}) \approx \hat{\mathbf{\Omega}}_g + (\hat{\mathbf{\Omega}}_g^{-1} \otimes \hat{\mathbf{\Omega}}_g^{-1} - \mu_{gg}\mathbf{I}_p \otimes \mathbf{I}_p)^{-1}(\mathbf{S}).$$

We also need an approximation for $\ln|\hat{\mathbf{\Omega}}_g(\mathbf{S})|$. By first-order Taylor expansion around $\mathbf{S} = \mathbf{0}$ we have

$$\begin{aligned} \ln|\hat{\mathbf{\Omega}}_g(\mathbf{S})| &\approx \ln|\hat{\mathbf{\Omega}}_g(\mathbf{0})| + \sum_{j,j'} \operatorname{tr}\Big[\hat{\mathbf{\Omega}}_g^{-1}(\mathbf{0})\frac{\partial\hat{\mathbf{\Omega}}_g}{\partial S_{jj'}}\Big]S_{jj'} \\ &\overset{\text{(S12)}}{\approx} \ln|\hat{\mathbf{\Omega}}_g(\mathbf{0})| + \operatorname{tr}\Big[\hat{\mathbf{\Omega}}_g^{-1}(\hat{\mathbf{\Omega}}_g \otimes \hat{\mathbf{\Omega}}_g + \mu_{gg}\hat{\mathbf{\Omega}}_g^2 \otimes \hat{\mathbf{\Omega}}_g^2)(\mathbf{S})\Big] \\ &= \ln|\hat{\mathbf{\Omega}}_g(\mathbf{0})| + \operatorname{tr}\big(\mathbf{S}\hat{\mathbf{\Omega}}_g + \mu_{gg}\hat{\mathbf{\Omega}}\mathbf{S}\hat{\mathbf{\Omega}}_g^2\big), \end{aligned} \tag{S13}$$

where we have used that $\frac{d}{dt}\ln|\mathbf{A}(t)| = \operatorname{tr}\big[\mathbf{A}(t)^{-1}\frac{d\mathbf{A}}{dt}\big]$ and $\frac{\partial \boldsymbol{\Omega}_g}{\partial S_{jj'}} \approx (\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg}\hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{E}_{jj'})$. We now have the necessary equations to derive the FKL approximate cross-validation score.

Define

$$f(\mathbf{A}, \mathbf{B}) = \ln|\mathbf{B}| - \operatorname{tr}(\mathbf{B}\mathbf{A}) \tag{S14}$$

by which the identity

$$\sum_{i=1}^{n_g} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g) = n_g f(\mathbf{S}_g, \boldsymbol{\Omega}_g) \tag{S15}$$

holds for all $g$. The full likelihood (3) in terms of $f$ is given by

$$\mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) \propto \sum_{g=1}^{G} \frac{n_g}{2}\Big\{\ln|\boldsymbol{\Omega}_g| - \operatorname{tr}(\boldsymbol{\Omega}_g\mathbf{S}_g)\Big\} = \sum_{g=1}^{G}\frac{n_g}{2} f(\mathbf{S}_g, \boldsymbol{\Omega}_g), \tag{S16}$$

while the likelihood of a single $\mathbf{S}_{ig}$ is

$$\mathcal{L}_{ig}(\boldsymbol{\Omega}_g; \mathbf{S}_{ig}) \propto \frac{1}{2}\Big\{\ln|\boldsymbol{\Omega}_g| - \operatorname{tr}(\boldsymbol{\Omega}_g\mathbf{S}_{ig})\Big\} = \frac{1}{2} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g). \tag{S17}$$

In our setting, the fused LOOCV score is given by:

$$\begin{aligned}
\text{LOOCV} &= -\frac{1}{n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathcal{L}_{ig}\big(\hat{\boldsymbol{\Omega}}_g^{\neg ig}; \mathbf{S}_{ig}\big) \\
&\overset{\text{(S17)}}{=} -\frac{1}{n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\frac{1}{2} f(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g^{\neg ig}) \\
&= -\frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{1}{2}\sum_{i=1}^{n_g}\Big[f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g\big) + f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g^{\neg ig}\big) - f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g\big)\Big] \\
&\overset{\text{(S15)}}{=} -\frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{n_g}{2} f\big(\mathbf{S}_g, \hat{\boldsymbol{\Omega}}_g\big) - \frac{1}{n_\bullet}\sum_{g=1}^{G}\frac{1}{2}\sum_{i=1}^{n_g}\Big[f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g^{\neg ig}\big) - f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g\big)\Big] \\
&\overset{\text{(S16)}}{=} -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}) - \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big[f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g^{\neg ig}\big) - f\big(\mathbf{S}_{ig}, \hat{\boldsymbol{\Omega}}_g\big)\Big] \\
&\overset{\text{(S14)}}{=} -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}) - \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big[\ln|\hat{\boldsymbol{\Omega}}_g^{\neg ig}| - \operatorname{tr}(\hat{\boldsymbol{\Omega}}_g^{\neg ig}\mathbf{S}_{ig}) - \ln|\hat{\boldsymbol{\Omega}}_g| + \operatorname{tr}(\hat{\boldsymbol{\Omega}}_g\mathbf{S}_{ig})\Big].
\end{aligned}$$

Now, substitution of (S12) and (S13) gives the FKL approximate cross-validation score as an approximation to the fused LOOCV score:

$$\text{LOOCV} \approx \widehat{\text{FKL}} = -\frac{1}{n_\bullet}\mathcal{L}(\{\hat{\boldsymbol{\Omega}}_g\}; \{\mathbf{S}_g\}) + \frac{1}{2n_\bullet}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\boldsymbol{\zeta}_{ig},$$

where

$$\begin{aligned}\boldsymbol{\zeta}_{ig} &= \operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}} + \mu_{gg}\hat{\boldsymbol{\Omega}}^2\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}} + \mu_{gg}\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) \\ &= \operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}\big) + \mu_{gg}\operatorname{tr}\big(\hat{\boldsymbol{\Omega}}^2\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}\big) - \mu_{gg}\operatorname{tr}\big(\hat{\boldsymbol{\Omega}}\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) \\ &= \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^2\big) + \mu_{gg}\operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^4\big) - \operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}\big) - \mu_{gg}\operatorname{tr}\big(\mathbf{S}\hat{\boldsymbol{\Omega}}^3\big) \\ &= \operatorname{tr}\big[\mathbf{S}(\hat{\boldsymbol{\Omega}}^2 - \hat{\boldsymbol{\Omega}})\big] + \mu_{gg}\operatorname{tr}\big[\mathbf{S}(\hat{\boldsymbol{\Omega}}^4 - \hat{\boldsymbol{\Omega}}^3)\big] \\ &= \mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}^2 - \hat{\boldsymbol{\Omega}})\mathbf{y}_{ig} + \mu_{gg}\mathbf{y}_{ig}^\top(\hat{\boldsymbol{\Omega}}^4 - \hat{\boldsymbol{\Omega}}^3)\mathbf{y}_{ig}. \end{aligned} \tag{S18}$$

To arrive at (S18) we have used the linear and cyclic properties of the trace operator. As $\mathbf{S} = \mathbf{y}_{ig}\mathbf{y}_{ig}^\top$, the cyclic property implies the final equality since $\operatorname{tr}(\mathbf{SA}) = \operatorname{tr}(\mathbf{y}_{ig}\mathbf{y}_{ig}^\top\mathbf{A}) = \operatorname{tr}(\mathbf{y}_{ig}^\top\mathbf{A}\mathbf{y}_{ig}) = \mathbf{y}_{ig}^\top\mathbf{A}\mathbf{y}_{ig}$. Equation (S18) is equivalent to the summand in (S7).

## 4. Additional Results Simulation Scenario 2

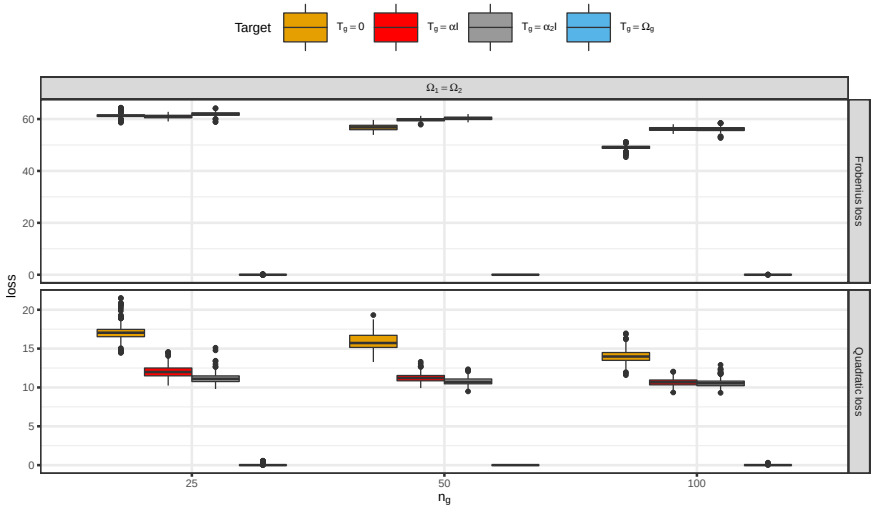


Figure S1: Results for simulation Scenario 2i. Comparison of the targeted versus the un-targeted approach in the banded population setting. We consider $G = 2$ classes with the population precision matrix $\boldsymbol{\Omega}$ for each class being a banded matrix with $p = 50$ and $k = 25$ bands. The considered class sample sizes are $n_g \in \{25, 50, 100\}$. The target matrix is taken to be equal over classes, i.e., $\mathbf{T}_1 = \mathbf{T}_2$. The un-targeted situation is represented by $\mathbf{T}_g = \mathbf{0}$. The most informative target is the spot-on target $\mathbf{T}_g = \boldsymbol{\Omega}$. Two diagonal targets are also considered: $\mathbf{T}_g = \alpha_\bullet\mathbf{I}_p$, with $\alpha_\bullet = [\sum_j(\mathbf{S}_\bullet)^{-1}_{jj}]/p$; and $\mathbf{T}_g = \alpha_{\bullet 2}\mathbf{I}_p$, with $\alpha_{\bullet 2} = p/\operatorname{tr}(\mathbf{S}_\bullet)$. Hence, $\alpha_\bullet$ represents the average of the inverse marginal variances of $\mathbf{S}_\bullet$ and $\alpha_{\bullet 2}$ represents the inverse of the averaged eigenvalues of $\mathbf{S}_\bullet$. Note that the boxplots in the figure (for each class sample size $n_g$) are ordered according to the legend (given at the top of the image).

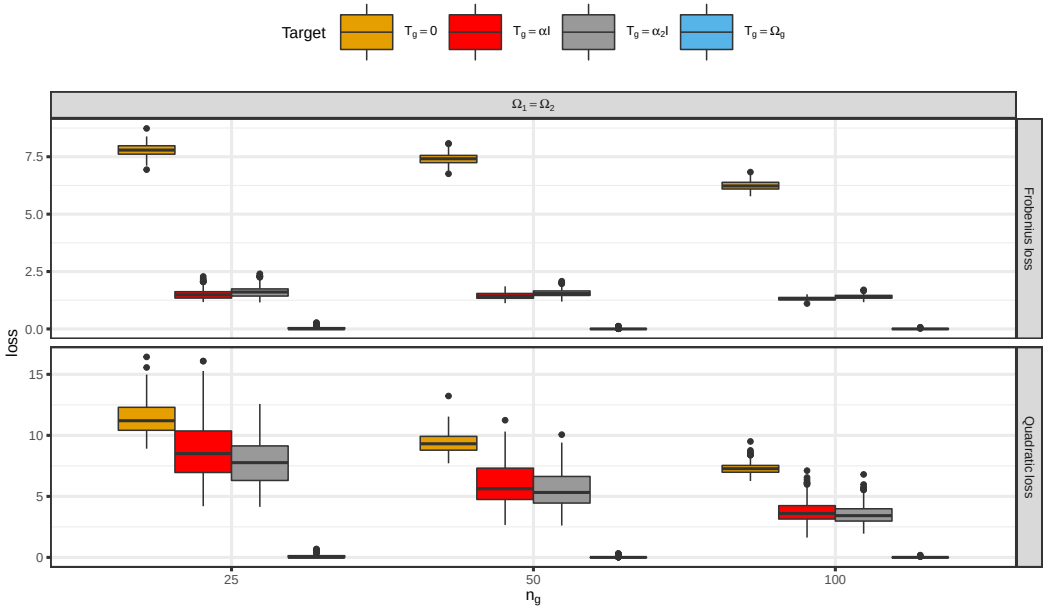


Figure S2: Results for simulation Scenario 2ii. Comparison of the targeted versus the un-targeted approach in the star population setting. We consider $G = 2$ classes with the population precision matrix $\mathbf{\Omega}$ for each class being a star matrix with $p = 50$ and where the first variable represents the internal node. The values of the off-diagonal entries $(1, j)$ and $(j, 1)$ taper-off by $1/(j + 1)$. The considered class sample sizes are $n_g \in \{25, 50, 100\}$. The target matrix is taken to be equal over classes, i.e., $\mathbf{T}_1 = \mathbf{T}_2$. The un-targeted situation is represented by $\mathbf{T}_g = \mathbf{0}$. The most informative target is the spot-on target $\mathbf{T}_g = \mathbf{\Omega}$. Two diagonal targets are also considered: $\mathbf{T}_g = \alpha_\bullet \mathbf{I}_p$, with $\alpha_\bullet = [\sum_j (\mathbf{S}_\bullet)^{-1}_{jj}]/p$; and $\mathbf{T}_g = \alpha_{\bullet 2} \mathbf{I}_p$, with $\alpha_{\bullet 2} = p/\mathrm{tr}(\mathbf{S}_\bullet)$. Hence, $\alpha_\bullet$ represents the average of the inverse marginal variances of $\mathbf{S}_\bullet$ and $\alpha_{\bullet 2}$ represents the inverse of the averaged eigenvalues of $\mathbf{S}_\bullet$. Note that the boxplots in the figure (for each class sample size $n_g$) are ordered according to the legend (given at the top of the image).

## 5. Additional Results Simulation Scenario 5

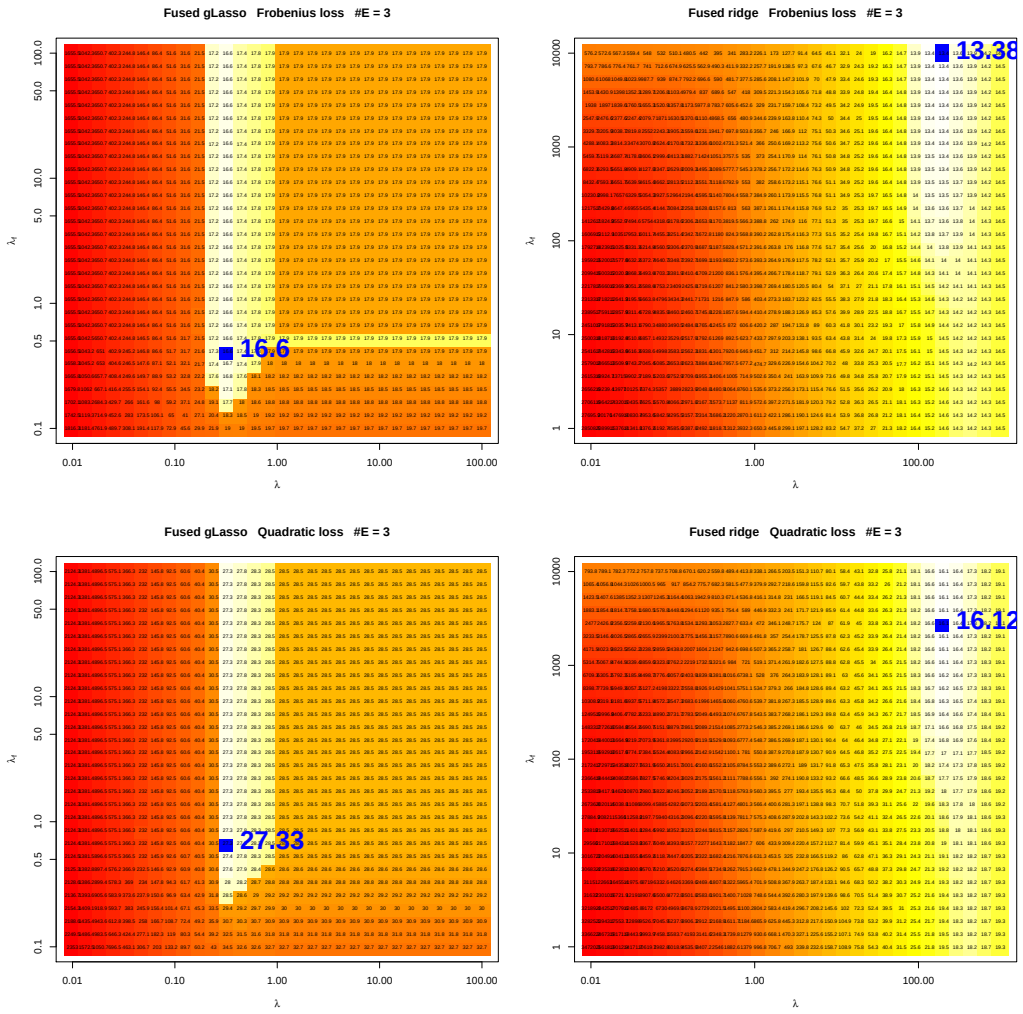


Figure S3: Comparison of the fused graphical lasso and the fused ridge estimator in the Barabási graph game population setting with $n_g = 25$ and where the number of edges to add in each time step was taken to be 3. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

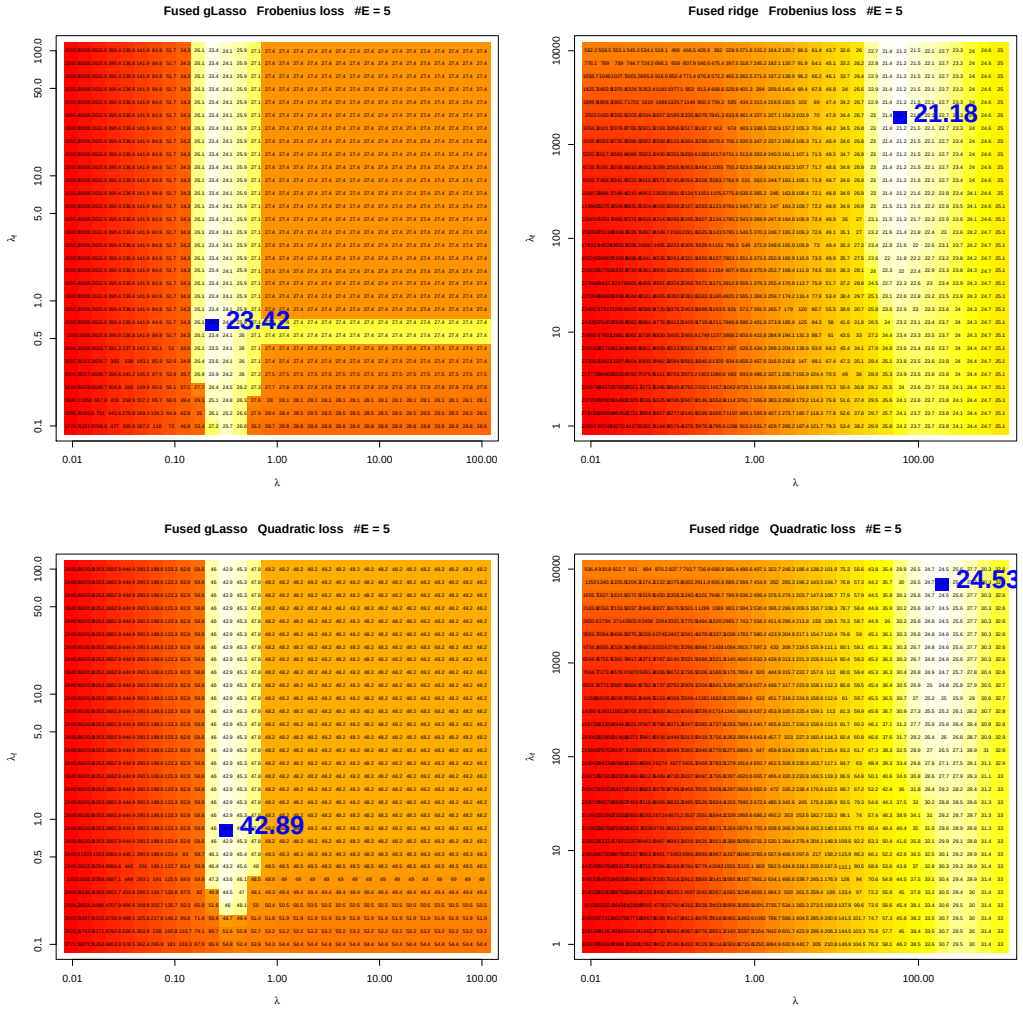


Figure S4: Comparison of the fused graphical lasso and the fused ridge estimator in the Barabási graph game population setting with $n_g = 25$ and where the number of edges to add in each time step was taken to be 5. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

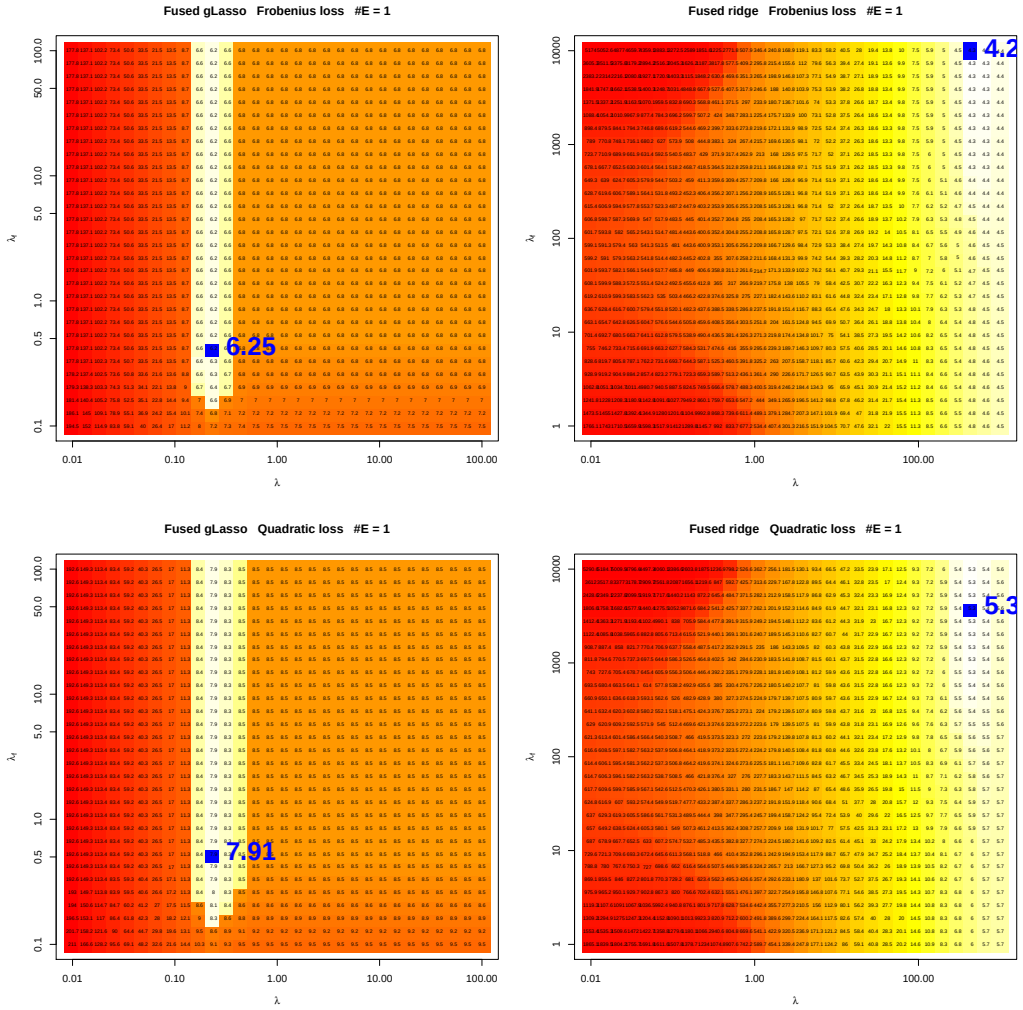


Figure S5: Comparison of the fused graphical lasso and the fused ridge estimator in the Barabási graph game population setting with $n_g = 50$ and where the number of edges to add in each time step was taken to be 1. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

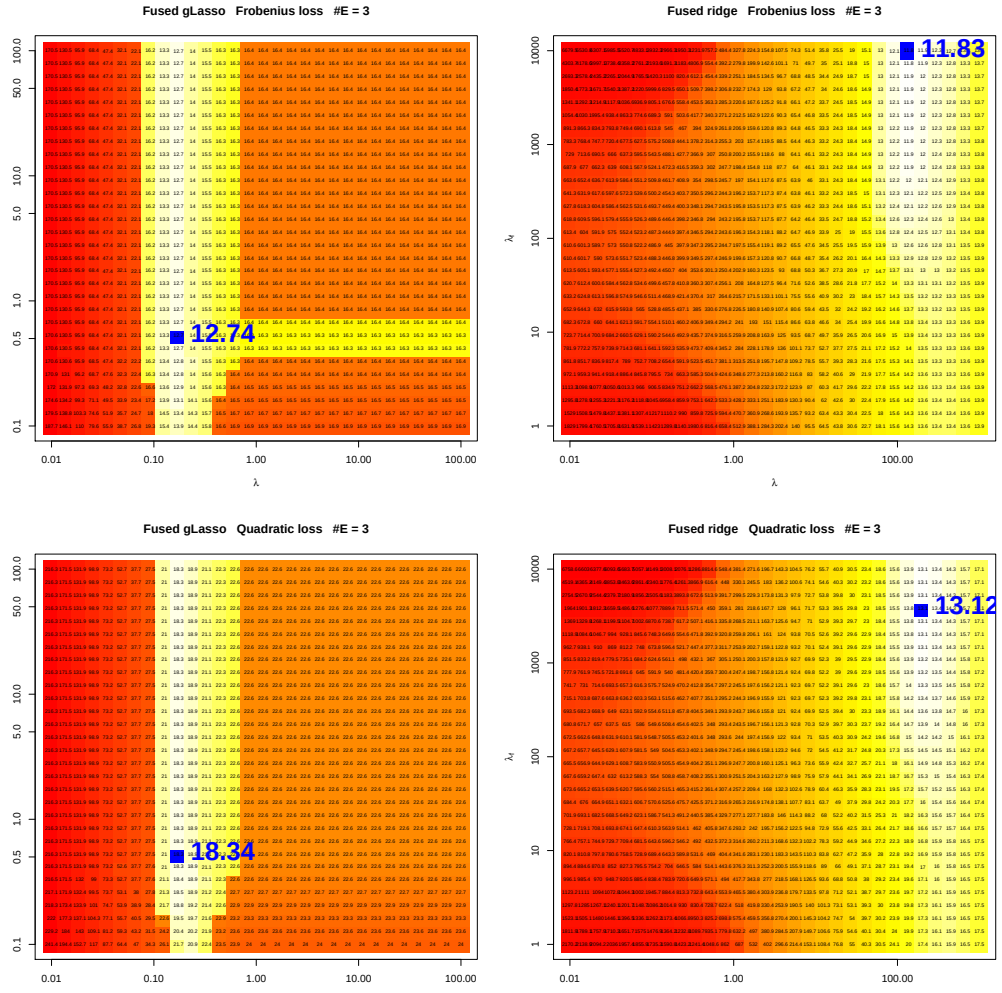


Figure S6: Comparison of the fused graphical lasso and the fused ridge estimator in the Barabási graph game population setting with $n_g = 50$ and where the number of edges to add in each time step was taken to be 3. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

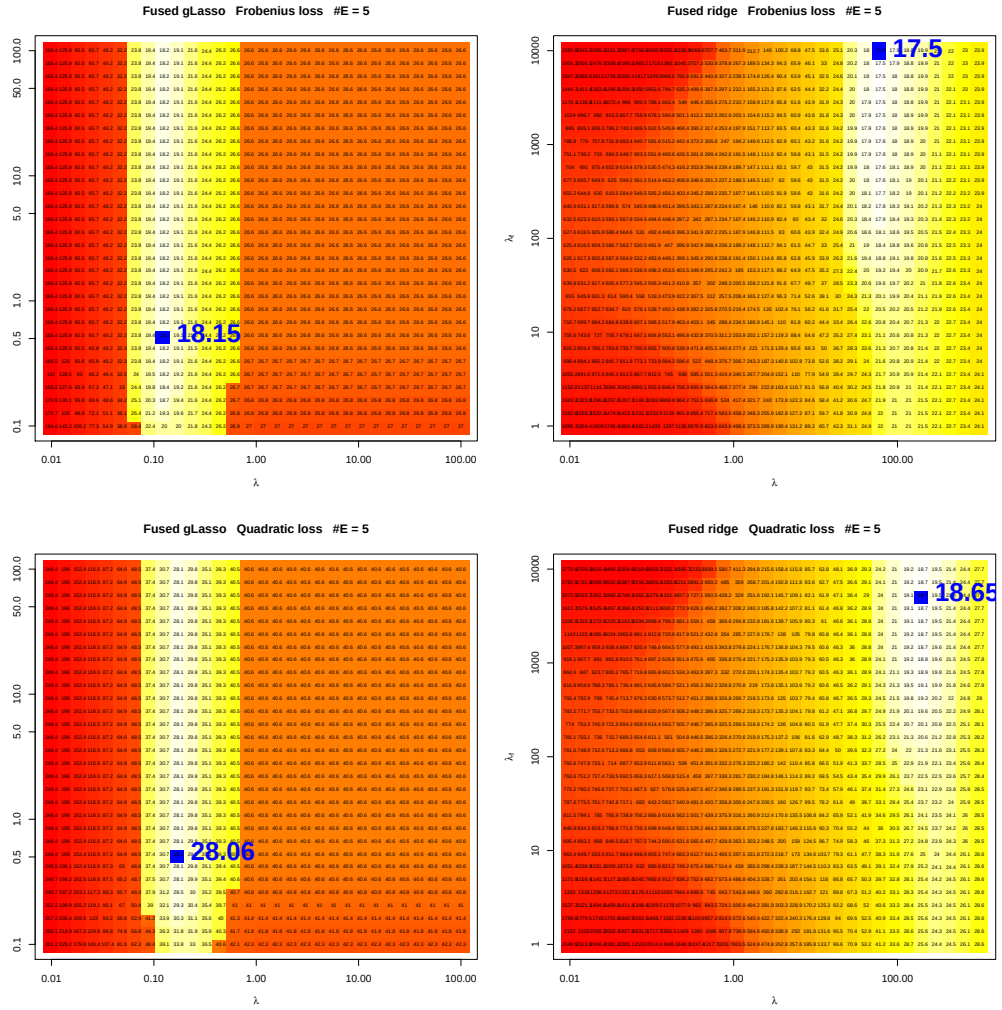


Figure S7: Comparison of the fused graphical lasso and the fused ridge estimator in the Barabási graph game population setting with $n_g = 50$ and where the number of edges to add in each time step was taken to be 5. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

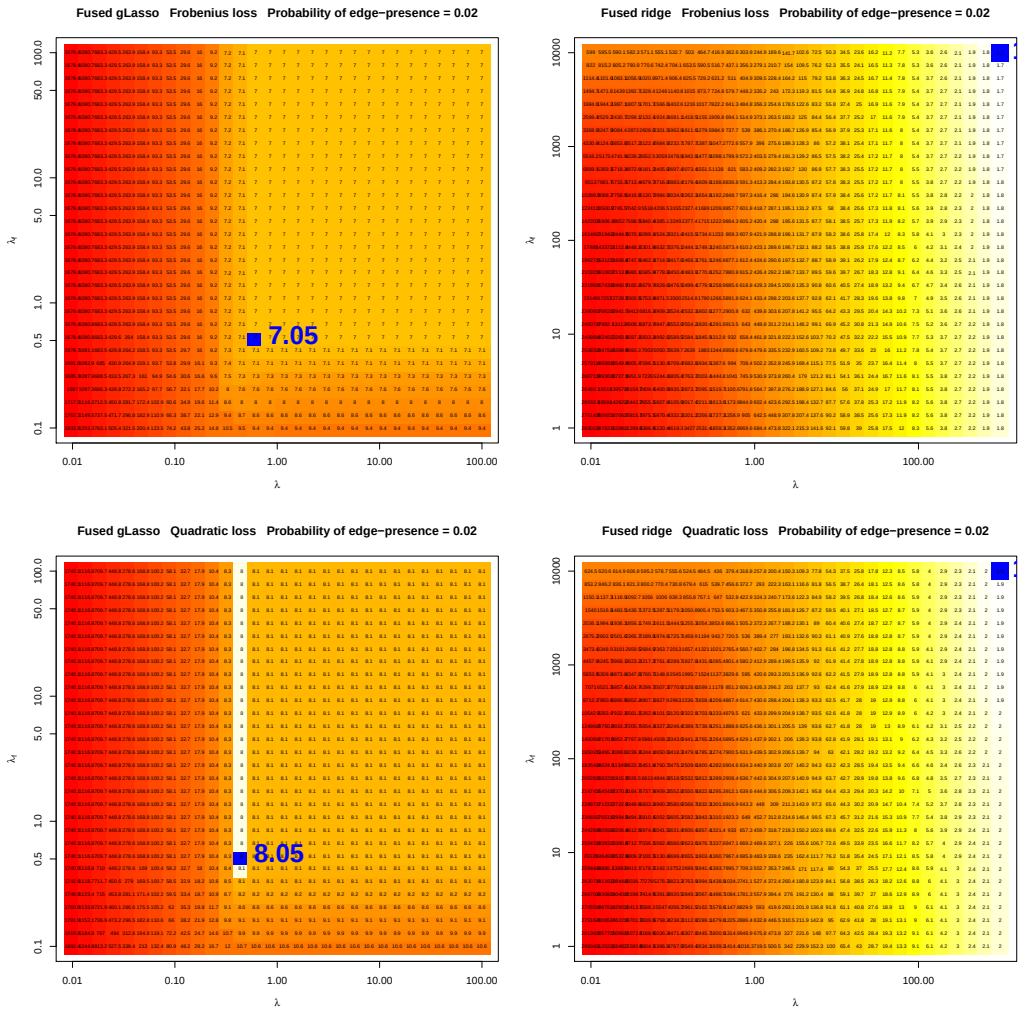


Figure S8: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 25$ and where the probability of edge-presence is set to $1/p = .02$. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

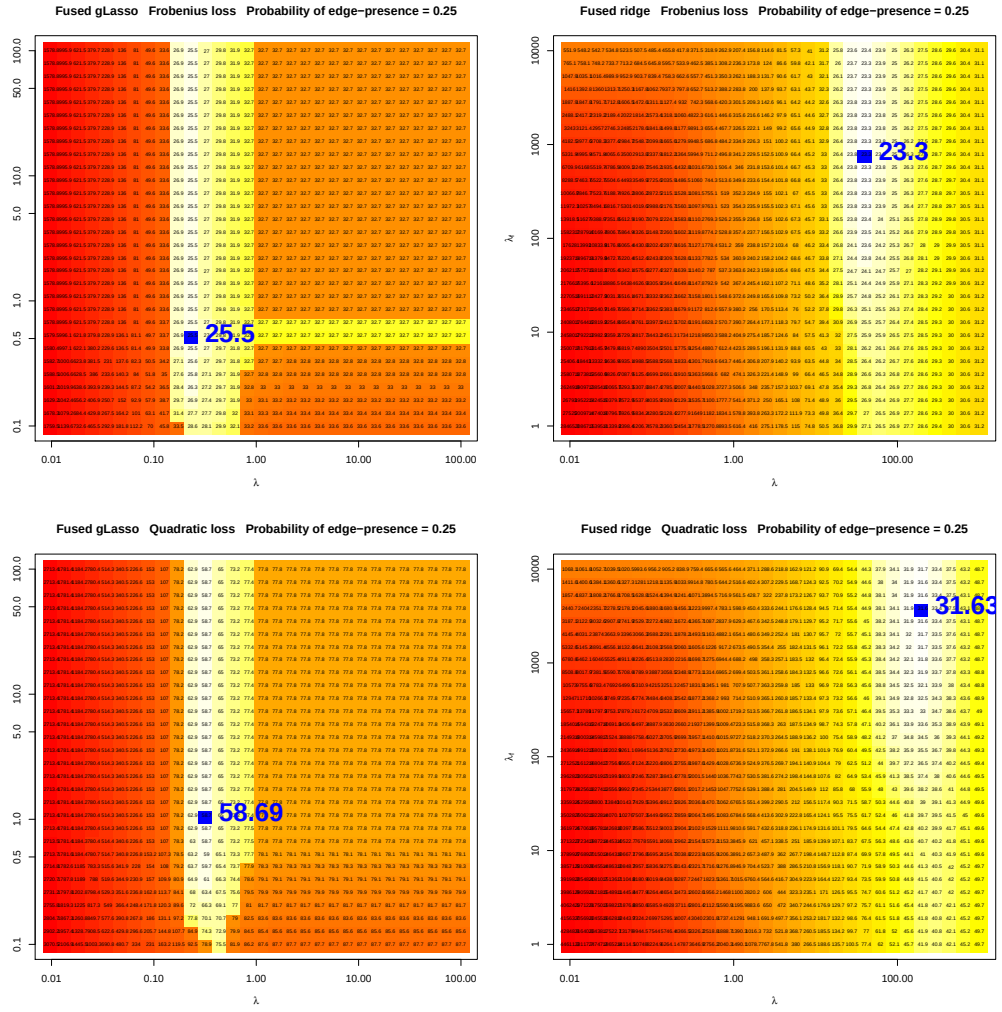


Figure S9: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 25$ and where the probability of edge-presence is set to .25. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

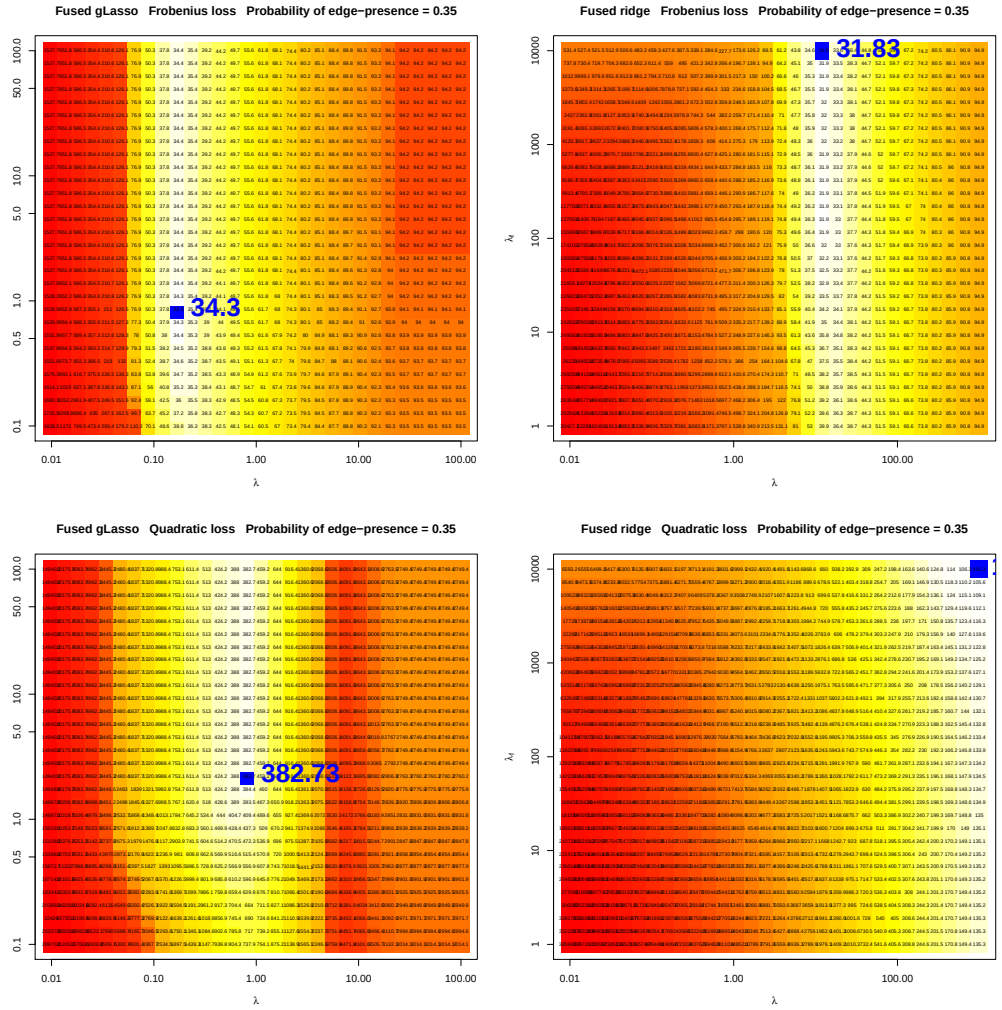


Figure S10: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 25$ and where the probability of edge-presence is set to .35. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

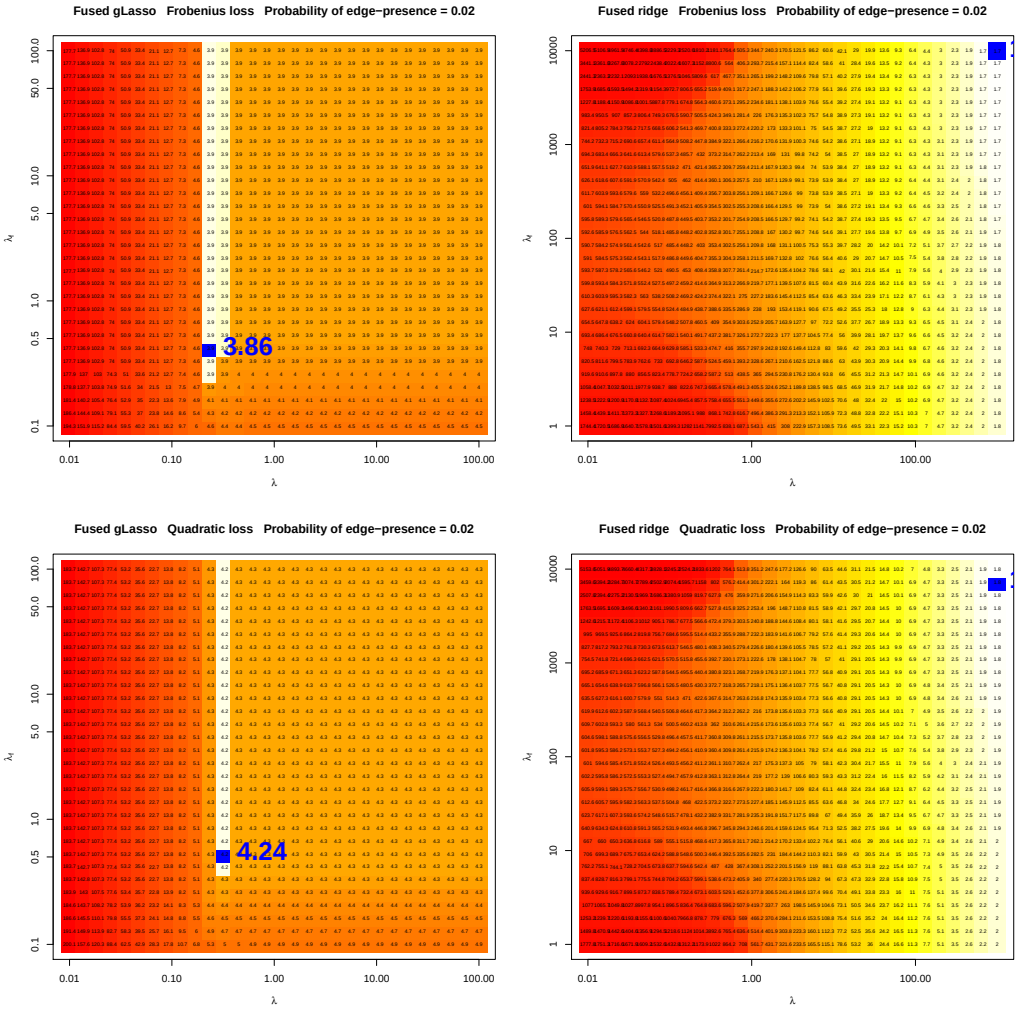


Figure S11: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 50$ and where the probability of edge-presence is set to $1/p = .02$. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

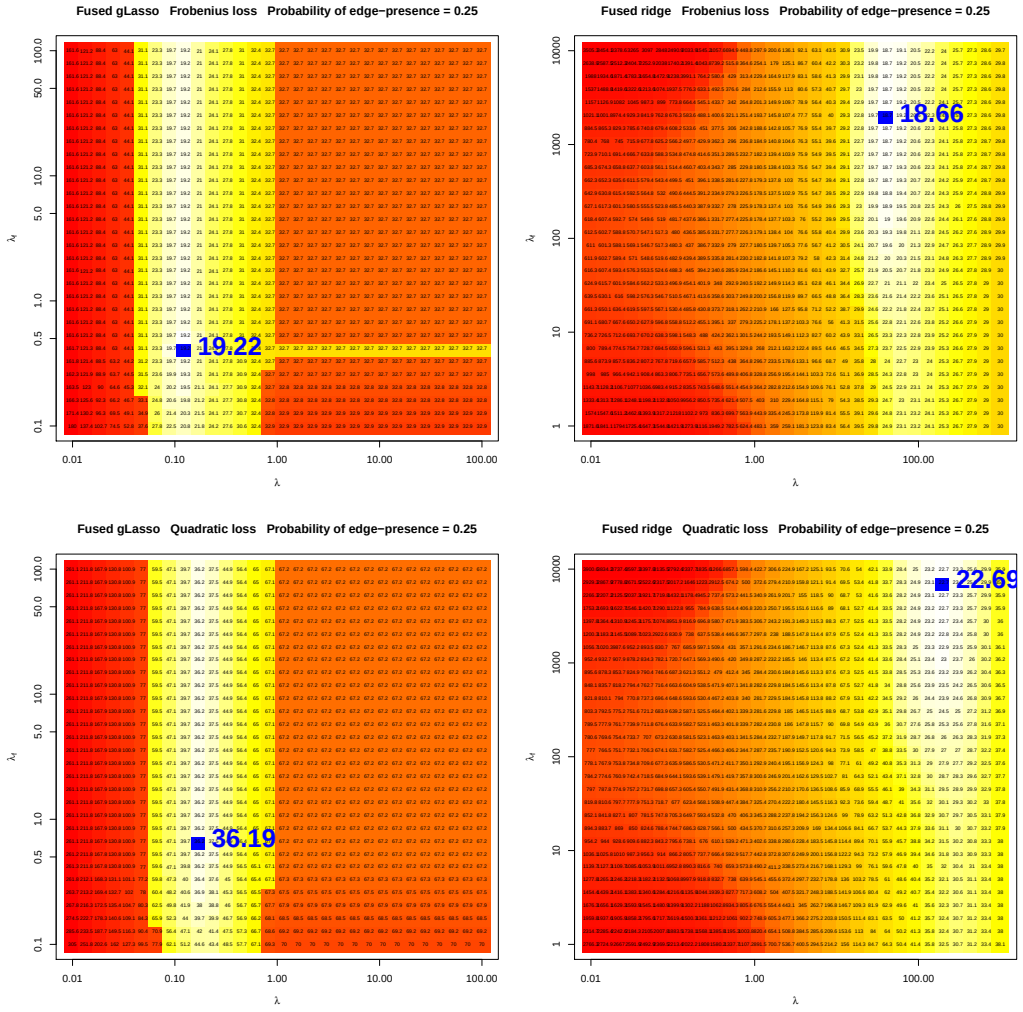


Figure S12: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 50$ and where the probability of edge-presence is set to .25. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

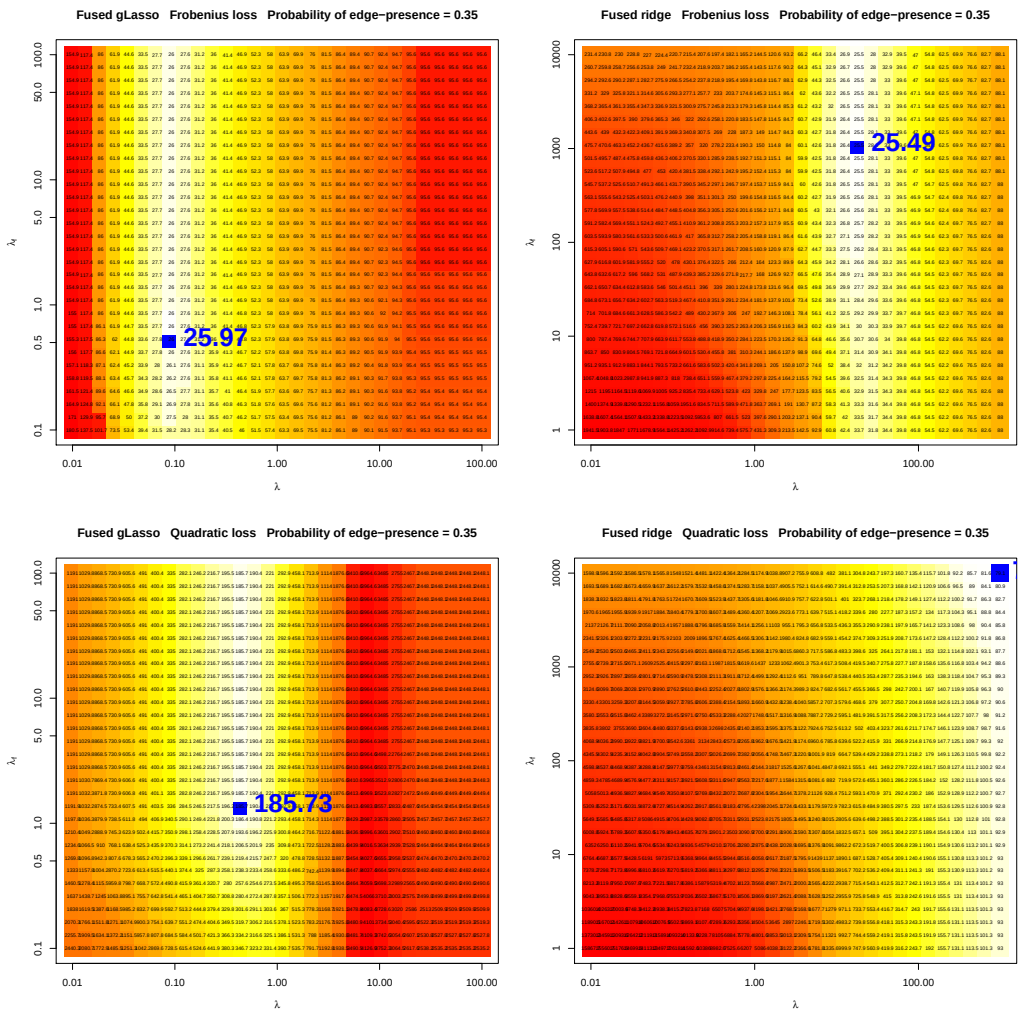


Figure S13: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 50$ and where the probability of edge-presence is set to .35. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

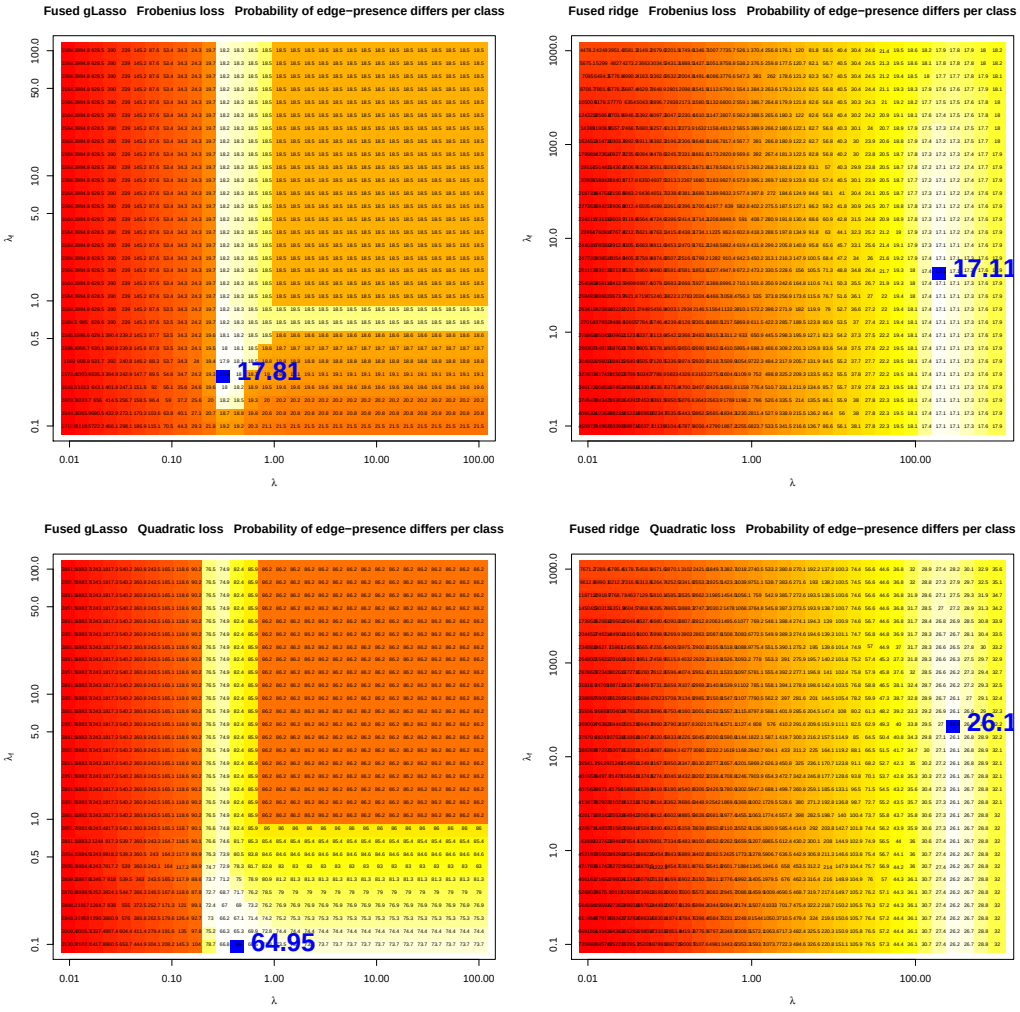


Figure S14: Comparison of the fused graphical lasso and the fused ridge estimator in the Erdös-Rényi random graph population setting with $n_g = 25$ under class dissimilarity. The probability of edge-presence is set to $1/p = .02$ for class 1 and .25 for class 2. Each square on the two-dimensional grid represents a $(\lambda, \lambda_f)$-combination. The number in each square represents the estimated Risk for the corresponding combination. The blue square (and corresponding number) indicate the lowest Risk achieved on the grid. Left-hand panels give the results for the fused graphical lasso. Right-hand panels give the results for the fused ridge estimator. Upper panels express the Risk surface under Frobenius loss. Lower panels express the Risk surface under quadratic loss.

## 6. Additional Results Simulation Scenario 6

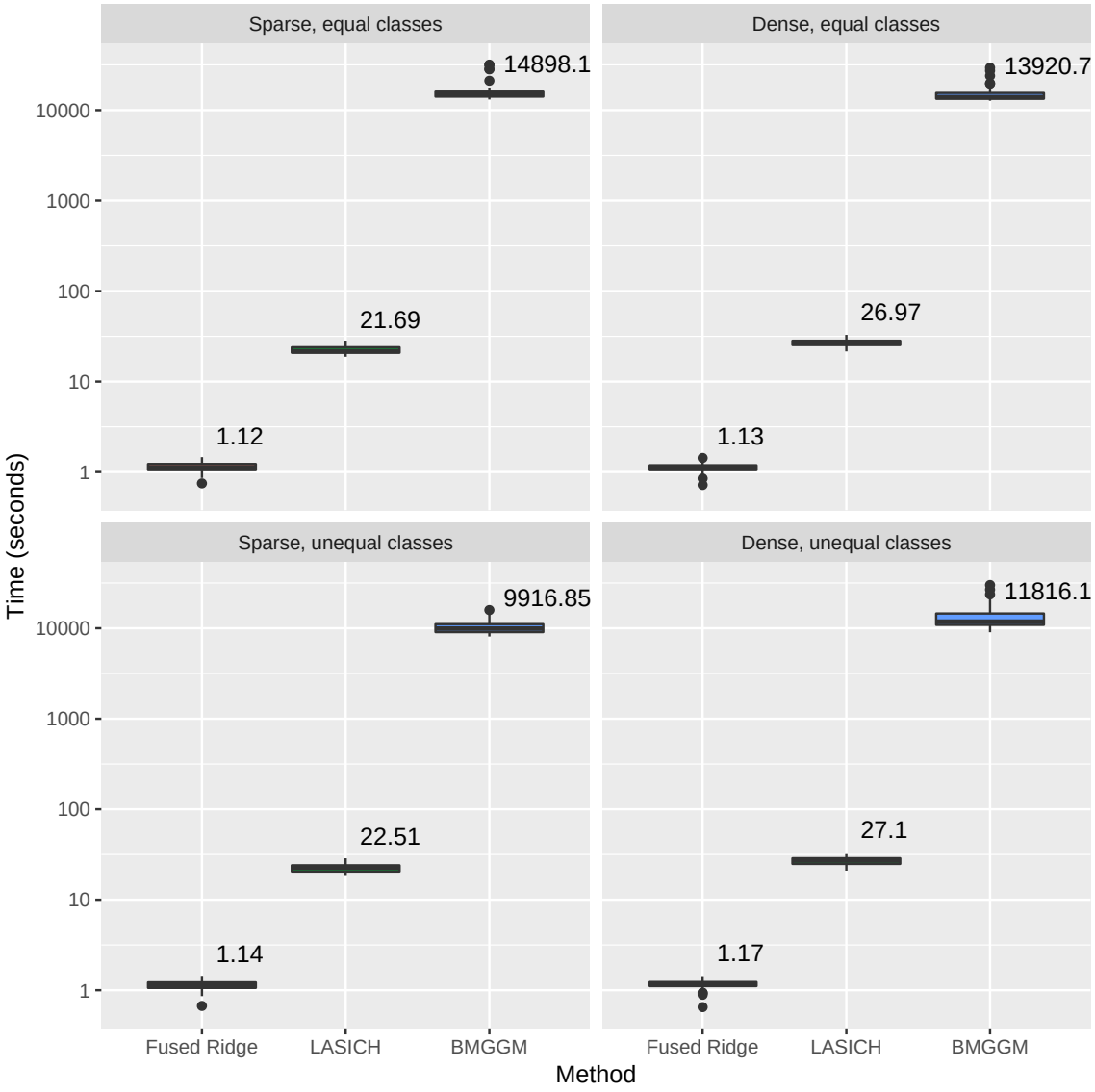


Figure S15: Timing results (in seconds) for the fused ridge, LASICH, and BMGGM methods for each of the considered sub-scenarios. The $x$-axis represents the methods. The $y$-axis has a logarithmic scale. Printed numbers above each boxplot then represent the median runtime for the respective method in a given sub-scenario.

## References

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc., 2010.

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

I. Vujačić, A. Abbruzzo, and E. Wit. A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *Journal of Statistical Computation and Simulation*, 85(18):3628–3640, 2015.