A EM Procedure for Multiple Means

We show that the Bayesian EM procedure provides sensible estimates of the regularization parameter even in the setting of the normal multiple means problem with known variance σ^2. In this setting, the LOOCV is unable to provide any guidance on how to choose λ due to all the information for each regression parameter being concentrated in a single observation. We use the τ parameterisation of the hyperparameter, rather than τ^2, as the resulting estimator has an easy to analyse form.

In the normal multiple means model, we are given $(y_i|\beta_i) \sim N(\beta_i, 1)$, i.e., $\mathbf{y}$ is a p-dimensional normally distributed vector with mean $\boldsymbol{\beta}$ and identity covariance matrix. The conditional posterior distribution of $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta}|\mathbf{y}, \tau \sim \mathrm{N}\left((1-\kappa)\mathrm{y}, \sigma^2(1-\kappa)\right) \tag{17}$$

where $\kappa = 1/(1+\tau^2)$. Under this setting, Strawderman [43] proved that if $p \geq 3$, then any estimator of the form

$$\left(1 - r\left(\frac{1}{2}||y||^2\right)\frac{p-2}{||y||^2}\right) y \tag{18}$$

where $0 \leq r\left(\frac{1}{2}||y||^2\right) \leq 2$ and $r(\cdot)$ is non-decreasing, is minimax, i.e., it dominates least-squares. We will now show that our EM procedure not only yields reasonable estimates in this setting, in contrast to LOOCV, but that these estimates are minimax, and hence dominate least-sqaures.

For the normal means model, we can obtain a closed form solution for the optimum τ, by solving for the stationary point for which $\tau_{t+1} = \tau_t$, with $\tau \sim C^+(0, 1)$:

$$\begin{aligned} \arg\min_{\tau} \{\mathrm{E}_{\boldsymbol{\beta}}[-\log\, p(y|\beta,\tau) - \log\, p(\beta|\tau) - \log\, \pi(\tau)]\} &= \tau \\ \arg\min_{\tau} \left\{\frac{p}{2}\log\tau^2 + \frac{w}{2\tau^2} + \log(1+\tau^2)\right\} &= \tau \\ \sqrt{\frac{w - p + \sqrt{p^2 + 8w + 2pw + w^2}}{2(2+p)}} &= \tau, \end{aligned}$$

and with $w = \sum_{j=1}^{p} \mathrm{E}\left[\beta_j^2\right] = (1-\kappa)^2 s + (1-\kappa)p$, $s = ||y||^2$ and $\tau = \sqrt{\frac{1-\kappa}{\kappa}}$. This yields

$$\sqrt{\frac{(\sqrt{p((\kappa-2)^2 p - 8\kappa + 8) - 2(\kappa-1)^2 s((\kappa-2)p-4) + (\kappa-1)^4 s^2} - \kappa p + (\kappa-1)^2 s)}{2(2+p)}} = \sqrt{\frac{1-\kappa}{\kappa}}$$

with solution $\kappa = (p+2)/s$. Plugging this κ solution into (17), we note that the resulting estimator of $\boldsymbol{\beta}$ (17) is of the form (18) with

$$\begin{aligned} r\left(\frac{1}{2}||y||^2\right) &= \left(\frac{p+2}{||y||^2}\right) / \left(\frac{p-2}{||y||^2}\right) \\ &= \frac{p+2}{p-2} \end{aligned}$$

As we have $r\left(\frac{1}{2}||y||^2\right) \leq 2$ when $p \geq 6$, the EM ridge estimator is minimax in this setting for $p \geq 6$.

B Proof of Theorem 3.1

We prove that for sufficiently large n, a continuous injective reparameterization of the negative log joint posterior of (5) & (7) is convex when restricted to $\tau^2 \geq \epsilon$. This is sufficient, since unimodality is preserved by strictly monotone transformations and continuous injective reparameterizations.

Specifically, for the presented hierarchical model, the negative log joint posterior up to an additive constant is

$$\frac{n+p+2}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{p+2-2a}{2}\log\tau^2 + \frac{\|\boldsymbol{\beta}\|^2}{2\sigma^2\tau^2} + (a+b)\log(1+\tau^2)$$

and reparameterising with $\phi = \beta/\sigma$, $\rho = 1/\sigma$ and $\chi = 1/\tau$ and reorganising terms yields

$$-(n+p+2)\log\rho - (p+2-2a)\log\chi + (a+b)\log(1+\chi^{-2}) + \frac{1}{2}\|\rho\mathbf{y} - \mathbf{X}\boldsymbol{\phi}\|^2 + \frac{\|\chi\boldsymbol{\phi}\|^2}{2} \ .$$

The first three terms can easily be checked to be convex via a second derivative test, for which the convexity of the second term is contingent on the condition that $a < 1 + p/2$, a condition that holds true in our specific scenario with $a = b = 1/2$. For the last two terms, the combined Hessian is of block form $[\mathbf{A}, \mathbf{B}; \mathbf{B}^{\mathrm{T}}, \mathbf{C}]$ with $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \chi^2\mathbf{I}_p$, $\mathbf{B} = 2\chi\boldsymbol{\phi}$, and $\mathbf{C} = \|\boldsymbol{\phi}\|^2$. Symmetric matrices of this form are positive definite if $\mathbf{A}$ and its Schur complement

$$\mathbf{C} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B} = \|\boldsymbol{\phi}\|^2 - 4\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/\chi^2 + \mathbf{I}_p)^{-1}\boldsymbol{\phi}$$

are positive definite. Clearly, $\mathbf{A}$ is positive definite. Moreover, for $n > 4/(\epsilon^2\gamma_n)$, we have

$$\begin{aligned}
\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/\chi^2 + I)^{-1}\boldsymbol{\phi} &= \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}/\chi^2 + I)^{-1}\boldsymbol{\phi} \\
&= \boldsymbol{\phi}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}/\chi^2 + I)^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\phi} \\
&\leq \boldsymbol{\phi}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}}\boldsymbol{\phi}\chi^2/(n\gamma_n) \\
&= ((\mathbf{I} - \mathbf{V}\mathbf{V}^{\mathrm{T}})\boldsymbol{\phi} + \mathbf{V}\mathbf{V}^{\mathrm{T}}\boldsymbol{\phi})^{\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}}\boldsymbol{\phi}\chi^2/(n\gamma_n) \\
&= \|\mathbf{V}\mathbf{V}^{\mathrm{T}}\boldsymbol{\phi}\|^2\chi^2/(n\gamma_n) \\
&\leq \|\boldsymbol{\phi}\|^2\chi^2/(n\gamma_n) \\
&< \|\boldsymbol{\phi}\|^2/(\epsilon^2 n\gamma_n) \\
&< \|\boldsymbol{\phi}\|^2/4
\end{aligned}$$

where we used the fact that $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ is the orthogonal projection onto the column space of $\mathbf{V}$. The overall inequality implies the required positivity of the Schur complement.

C Derivation of Equation 9

C.1 Derivation of ESN

Here we show that

$$\begin{aligned}
\sum_{j=1}^{p} \mathrm{E}\left[\beta_j^2 \mid \hat{\tau}^{(t)}, \hat{\sigma}^{2^{(t)}}\right] &= \mathrm{tr}\left(\mathrm{Cov}[\boldsymbol{\beta}]\right) + \sum_{j=1}^{p} \mathrm{E}[\beta_j]^2 \\
&= \sigma^2\mathrm{tr}(\mathbf{A}_\tau^{-1}) + \|\hat{\boldsymbol{\beta}}_\tau\|^2
\end{aligned}$$

This is a rather straightforward proof. We use the fact that given a random variable x; the expected squared value of x is $\mathrm{E}\left[x^2\right] = \mathrm{Var}[x] + \mathrm{E}[x]^2$.

C.2 Derivation of ESS

Here we show that

$$\mathrm{E}_{\boldsymbol{\beta}}\left[||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \mid \hat{\tau}^{(t)}, \hat{\sigma}^{2^{(t)}}\right] = ||\mathbf{y} - \mathbf{X}\ \mathrm{E}[\boldsymbol{\beta}]\ ||^2 + \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\ \mathrm{Cov}[\boldsymbol{\beta}]) \tag{19}$$

We first provide an important fact on the quadratic forms of random variables in Lemma C.1 below:

Lemma C.1. *Let* $\mathbf{b}$ *be a p-dimensional random vector and* $\mathbf{A}$ *be a p-dimensional symmetric matrix. If* $\mathrm{E}[\mathbf{b}] = \boldsymbol{\mu}$ *and* $\mathrm{Var}(\mathbf{b}) = \boldsymbol{\Sigma}$*, then* $\mathrm{E}\left[\mathbf{b}^{\mathrm{T}}\mathbf{A}\mathbf{b}\right] = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}$.

Now, we expand the left-hand side of Equation 19 :

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\beta}}\left[||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2\right] &= \mathrm{E}_{\boldsymbol{\beta}}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\,\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathrm{E}[\boldsymbol{\beta}] + \mathrm{E}\left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\right]
\end{aligned} \tag{20}$$

The use of lemma C.1 allows Equation 20 to be rewritten as

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\beta}}\left[||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2\right] &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\,\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathrm{E}[\boldsymbol{\beta}] + \mathrm{E}[\boldsymbol{\beta}]^{\mathrm{T}}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathrm{E}[\boldsymbol{\beta}] + \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\ \mathrm{Cov}[\boldsymbol{\beta}]) \\
&= ||\mathbf{y} - \mathbf{X}\ \mathrm{E}[\boldsymbol{\beta}]\ ||^2 + \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\ \mathrm{Cov}[\boldsymbol{\beta}]) \\
&= ||\mathbf{y} - \mathbf{X}\ \hat{\boldsymbol{\beta}}_\tau||^2 + \sigma^2\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}_\tau^{-1})
\end{aligned}$$

D Solving for the parameter updates (Derivation of Equation 11)

Rather than solving a two-dimensional numerical optimization problem (10), we show that given a fixed τ^2, we can find a closed formed solution for σ^2, and vice versa. To start off, we need to find the solution for σ^2 as a function of τ^2. First, find the negative logarithm of the joint probability distribution of hierarchy (5):

$$\arg\min_{\sigma^2} \left\{ \mathrm{E}_{\boldsymbol{\beta}}\left[-\log\ p(y|X,\beta,\sigma^2) - \log\ p(\beta|\tau^2,\sigma^2) - \log\ p(\sigma^2) - \log\ \pi(\tau^2) \right] \right\}. \tag{21}$$

Dropping terms that do not depend on σ^2 yields:

$$\begin{aligned} &\arg\min_{\sigma^2} \left\{ \mathrm{E}_{\boldsymbol{\beta}}\left[-\log\ p(y|X,\beta,\sigma^2) - \log\ p(\beta|\tau^2,\sigma^2) - \log\ p(\sigma^2) \right] \right\} \\ &= \arg\min_{\sigma^2} \left\{ \left(\frac{n+p}{2} \right) \log\sigma^2 + \frac{\mathrm{ESS}}{2\sigma^2} + \frac{\mathrm{ESN}}{2\sigma^2\tau^2} + \log\sigma^2 \right\} \\ &= \arg\min_{\sigma^2} \left\{ \left(\frac{n+p+2}{2} \right) \log\sigma^2 + \frac{\mathrm{ESS}}{2\sigma^2} + \frac{\mathrm{ESN}}{2\sigma^2\tau^2} \right\}. \end{aligned} \tag{22}$$

Solving the above minimization problem involves differentiating the negative logarithm with respect to σ^2 and solving for σ^2 that set the derivative to zero. This gives us:

$$\begin{aligned} \frac{\partial}{\partial\sigma^2} \left\{ \left(\frac{n+p+2}{2} \right) \log\sigma^2 + \frac{\mathrm{ESS}}{2\sigma^2} + \frac{\mathrm{ESN}}{2\sigma^2\tau^2} \right\} &= 0 \\ \frac{2+n+p}{2\sigma^2} - \frac{\mathrm{ESS}}{2(\sigma^2)^2} - \frac{\mathrm{ESN}}{2(\sigma^2)^2\tau^2} &= 0 \\ \hat{\sigma}^2 &= \frac{\tau^2\mathrm{ESS} + \mathrm{ESN}}{(n+p+2)\tau^2} \end{aligned} \tag{23}$$

Next, to obtain the M-step updates for the shrinkage parameter τ^2, we repeat the same procedure - find the negative logarithm of the joint probability distribution and remove terms that do not depend on either σ^2 or τ^2:

$$\begin{aligned} &\arg\min_{\tau^2} \left\{ \mathrm{E}_{\boldsymbol{\beta}}\left[-\log\ p(y|X,\beta,\sigma^2) - \log\ p(\beta|\tau^2,\sigma^2) - \log\ p(\sigma^2) - \log\ \pi(\tau^2) \right] \right\} \\ &= \arg\min_{\tau^2} \left\{ \left(\frac{n+p+2}{2} \right) \log\sigma^2 + \frac{\mathrm{ESS}}{2\sigma^2} + \frac{\mathrm{ESN}}{2\sigma^2\tau^2} + \frac{p}{2}\log\tau^2 + \log(1+\tau^2) + \frac{\log\tau^2}{2} \right\} \end{aligned} \tag{24}$$

Substiting the solution for σ^2 (23) into equation (24), yields a Q-function that depends only on τ^2. We eliminate the dependency on σ^2 by finding the optimal σ^2 as a function of τ^2 and substitute it into the Q-function of (24):

$$\arg\min_{\tau^2} \left\{ \frac{1}{2} \left[(1+p)\log\tau^2 + 2\log(1+\tau^2) + (n+p+2) \left(1 + \log\left(\frac{\mathrm{ESN} + \tau^2\mathrm{ESS}}{(n+p+2)\tau^2} \right) \right) \right] \right\} \tag{25}$$

Differentiating (25) with respect to τ^2 and solving for the τ^2 that set the derivative to zero yields:

$$\begin{aligned} \frac{\partial}{\partial\tau^2} \left\{ \frac{1}{2} \left[(1+p)\log\tau^2 + 2\log(1+\tau^2) + (n+p+2) \left(1 + \log\left(\frac{\mathrm{ESN} + \tau^2\mathrm{ESS}}{(n+p+2)\tau^2} \right) \right) \right] \right\} &= 0 \\ \frac{(3\mathrm{ESS} + \mathrm{ESS}p)(\tau^2)^2 + (\mathrm{ESN} - \mathrm{ESN}n + \mathrm{ESS} + \mathrm{ESS}p)\tau^2 - \mathrm{ESN} - \mathrm{ESN}n}{2\tau^2(1+\tau^2)(\mathrm{ESN} + \tau^2\mathrm{ESS})} &= 0. \end{aligned} \tag{26}$$

The τ^2 update is the positive solution to the quadratic equation (in terms of τ^2) (26):

$$\hat{\tau}^2 = \frac{(n-1)\mathrm{ESN} - (1+p)\mathrm{ESS} + \sqrt{(4n+4)\mathrm{ESN}(3+p)\mathrm{ESS} + ((1-n)\mathrm{ESN} + (p+1)\mathrm{ESS})^2}}{(6+2p)\mathrm{ESS}}$$

Table 3: Pseudo-code of EM algorithm with complexity of individual steps.

EM Algorithm with SVD	Operations
Input: Standardised predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, centered targets $\mathbf{y} \in \mathbb{R}^n$ and convergence threshold $\epsilon > 0$	
Output: $\boldsymbol{\beta} \in \mathbb{R}^p$	
$r = \min(n, p)$	$O(1)$
IF $p \geq n$	
$[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \mathrm{svd}(\mathbf{X})$	$O(mr^2)$
$\mathbf{s}^2 = (\Sigma_{1,1}^2, \ldots, \Sigma_{r,r}^2)$	$O(r)$
$\mathbf{c} = (\mathbf{U}^{\mathrm{T}}\mathbf{y}) \odot \mathbf{s}$	$O(nr)$
ELSE	
$[\mathbf{V}, \mathbf{\Sigma}^2] = \mathrm{eigen}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$	$O(mr^2)$
$\mathbf{s}^2 = (\mathbf{\Sigma}_{1,1}^2, \ldots, \mathbf{\Sigma}_{r,r}^2)$	$O(r)$
$\mathbf{c} = \mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$	$O(np)$
$Y = \mathbf{y}^{\mathrm{T}}\mathbf{y}$	$O(n)$
$\tau^2 \leftarrow 1$	$O(1)$
$\sigma^2 \leftarrow (1/n)\sum_{i=1}^n (y_i - \bar{y})^2,\ \bar{y} = (1/n)\sum_{i=1}^n y_i$	$O(n)$
$\mathrm{RSS} \leftarrow \infty$	$O(1)$
DO	
$\mathrm{RSS}_{\mathrm{old}} \leftarrow \mathrm{RSS}$	$O(k)$
$\alpha_j \leftarrow \dfrac{c_j}{s_j^2 + 1/\tau^2}$	$O(kr)$
(E-step)	
$\mathrm{ESN} \leftarrow \sum_{j=1}^r \alpha_j^2 + \sigma^2 \left(\sum_{j=1}^r \dfrac{1}{s_j^2 + \tau^{-2}} + \tau^2 \max(p-n, 0) \right)$	$O(kr)$
$\mathrm{RSS} \leftarrow Y - 2\sum_{j=1}^r \alpha_j c_j + \sum_{j=1}^r \alpha_j^2 s_j^2$	$O(kr)$
$\mathrm{ESS} \leftarrow \mathrm{RSS} + \sigma^2 \left(\sum_{j=1}^r \dfrac{s_j^2}{s_j^2 + \tau^{-2}} \right)$	$O(kr)$
(M-step)	
$g \leftarrow (4n+4)\mathrm{ESN}\,(3+p)\mathrm{ESS} + ((1-n)\mathrm{ESN} + (p+1)\mathrm{ESS})^2$	$O(k)$
$\tau^2 \leftarrow \dfrac{(n-1)\mathrm{ESN} - (1+p)\mathrm{ESS} + \sqrt{g}}{(6+2p)\mathrm{ESS}}$	$O(k)$
$\sigma^2 \leftarrow \dfrac{\tau^2\mathrm{ESS} + \mathrm{ESN}}{(n+p+2)\tau^2}$	$O(k)$
$\delta \leftarrow \dfrac{\lvert \mathrm{RSS}_{\mathrm{old}} - \mathrm{RSS} \rvert}{(1 + \lvert \mathrm{RSS} \rvert)}$	$O(k)$
until $\delta < \epsilon$	
$\alpha_j \leftarrow \dfrac{c_j}{s_j^2 + 1/\tau^2}$	$O(kr)$
$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\alpha}$	$O(pr)$
return $\boldsymbol{\beta}$	

Table 4: Pseudocode of the fast LOOCV algorithm with complexity of individual steps. $\mathbf{R}$ has column vectors $\mathbf{r}_j$ for $1 \leq j \leq r$.

Fast LOOCV ridge with SVD	Operation
Input: Standardised predictors $\mathbf{X} \in \mathbb{R}^{n\times p}$, centered targets $\mathbf{y} \in \mathbb{R}^n$ and a grid of penalty parameters $L = (\lambda_1, \lambda_2, \ldots, \lambda_l)$	
Output: $\boldsymbol{\beta} \in \mathbb{R}^p$	
$r = \min(n, p)$	$O(1)$
IF $p \geq n$	
$[\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}] = \mathrm{svd}(\mathbf{X})$	$O(mr^2)$
$\mathbf{s} = (\Sigma_{1,1}, \ldots, \Sigma_{r,r})$	$O(r)$
$\mathbf{R} = (s_1\mathbf{u}_1, \ldots, s_r\mathbf{u}_r)$	$O(nr)$
$\mathbf{c} = (\mathbf{U}^{\mathrm{T}}\mathbf{y}) \odot \mathbf{s}$	$O(nr)$
ELSE	
$[\mathbf{V}, \boldsymbol{\Sigma}^2] = \mathrm{eigen}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$	$O(mr^2)$
$\mathbf{s}^2 = (\boldsymbol{\Sigma}^2_{1,1}, \ldots, \boldsymbol{\Sigma}^2_{r,r})$	$O(r)$
$\mathbf{R} = \mathbf{X}\mathbf{V}$	$O(nrp)$
$\mathbf{c} = \mathbf{R}^{\mathrm{T}}\mathbf{y}$	$O(nr)$
$\mathbf{U} = (\mathbf{r}_1/s_1, \ldots, \mathbf{r}_r/s_r)$	$O(nr)$
for $\lambda \in L$ {	
$h_i = \sum_{j=1}^{r} \left(\frac{s_j^2}{s_j^2 + \lambda}\right) u_{ij}^2, \qquad (i = 1, \ldots, n)$	$O(lnr)$
$\alpha_j = \frac{c_j}{s_j^2 + \lambda}$	$O(lr)$
$\mathbf{e} = \mathbf{y} - \mathbf{R}\boldsymbol{\alpha}$	$O(lnr)$
$\mathrm{CVE}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1 - h_i}\right)^2$	$O(ln)$
}	
Find $\lambda^* = \arg\min_{\lambda \in L} \{\mathrm{CVE}(\lambda)\}$	$O(l)$
$\alpha_j = \frac{c_j}{s_j^2 + \lambda^*}$	$O(r)$
$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\alpha}$	$O(pr)$
return $\boldsymbol{\beta}$	

E Supplementary Results Material

E.1 Real Datasets Details

Table 5: Real datasets details

DATASETS	ABBREVIATION	n	p	TARGET VARIABLE	SOURCE
BUZZ IN SOCIAL MEDIA (TWITTER)	TWITTER	583250	77	mean number of active discussion	UCI
BLOG FEEDBACK	BLOG	60021	281	number of comments in the next 24 hours	UCI
RELATIVE LOCATION OF CT SLICES ON AXIAL AXIS	CT SLICES	53500	386	reference: Relative image location on axial axis	UCI
BUZZ IN SOCIAL MEDIA (TOM'S HARDWARE)	TOMSHW	28179	97	Mean Number of display	UCI
CONDITION-BASED MAINTENANCE OF NAVAL PROPULSION PLANTS	NPD - COM	11934	16	GT Compressor decay state coefficient	UCI
CONDITION-BASED MAINTENANCE OF NAVAL PROPULSION PLANTS	NPD - TUR	11934	16	GT Turbine decay state coefficient	UCI
PARKINSON'S TELEMONITORING	PT - MOTOR	5875	26	motor UPDRS score	UCI
PARKINSON'S TELEMONITORING	PT - TOTAL	5875	26	total UPDRS score	UCI
ABALONE	ABALONE	4177	8	Rings (age in years)	UCI
COMMUNITIES AND CRIME	CRIME	1994	128	ViolentCrimesPerPop	UCI
AIRFOIL SELF-NOISE	AIRFOIL	1503	6	Scaled sound pressure level (decibels)	UCI
STUDENT PERFORMANCE	STUDENT	649	33	final grade (with G1 & G2 removed)	UCI
CONCRETE COMPRESSIVE STRENGTH	CONCRETE	1030	9	Concrete compressive strength (MPa)	UCI
FOREST FIRES	F.FIRES	517	13	forest burned area (in ha)	UCI
BOSTON HOUSING	B.HOUSING	506	13	Median value of owner-occupied homes in $1000's	[21]
FACEBOOK METRICS	FACEBOOK	500	19	Total Interactions (with comment, like, and share columns removed)	UCI
DIABETES	DIABETES	442	10	quantitative measure of disease progression one year after baseline	[13]
REAL ESTATE VALUATION	R.ESTATE	414	7	house price of unit area	UCI
AUTO MPG	A.MPG	398	8	city-cycle fuel consumption in miles per gallon	UCI
YACHT HYDRODYNAMICS	YACHT	308	7	residuary resistance per unit weight of displacement	UCI
AUTOMOBILE	A.MOBILE	205	26	price	UCI
RAT EYE TISSUES	EYE	120	200	the expression level of TRIM32 gene	[39]
RIBOFLAVIN	RIBO	71	4088	Log-transformed riboflavin production rate	[9]
CROP	CROP	24000	3072	24 crop classes	UCR
ELECTRIC DEVICES	ELECD	16637	4096	7 electric devices	UCR
STARLIGHT CURVES	STARL	9236	7168	3 starlight curves	UCR

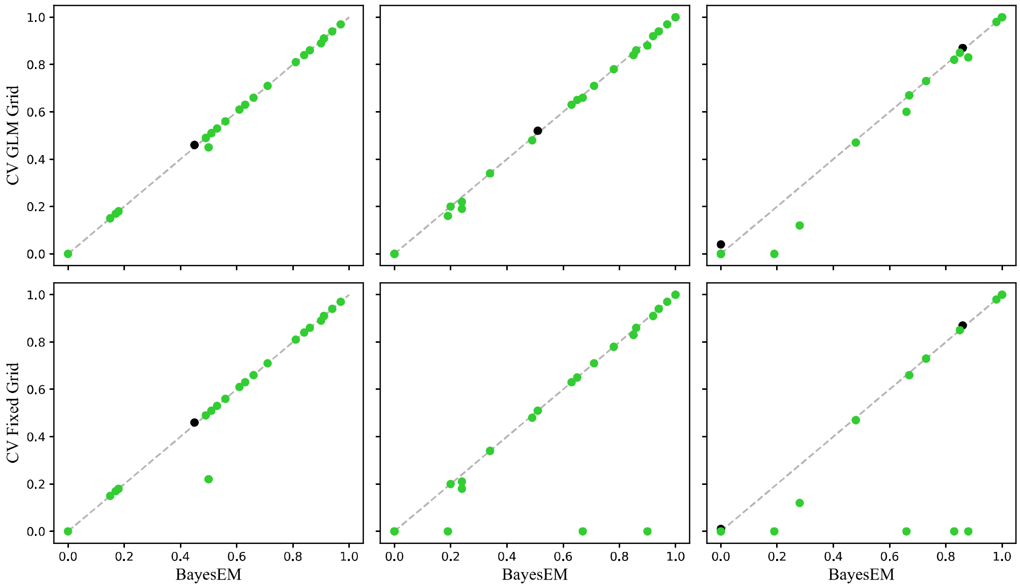

Figure 3: Comparison of predictive performance (R^2) of EM algorithm (x-axes) against CV with fixed grid (y-axes, top) and `glmnet` heuristic (y-axis, bottom). Columns correspond to the results of linear features (left), second-order features (middle), and third-order features (right). Negative values are capped at 0. Points skewing toward the bottom right indicate when our EM approach is giving better/same prediction performance as LOOCV (colored in green).