To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
August 15, 2017

Chapter 3

Computational methods for I-priors

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Theorem 3.1 (Euclid)**.** *For every prime p, there is a prime $p' > p$. In particular, the list of primes,*

$$2, 3, 5, 7, \ldots \tag{3.1}$$

is infinite.

Proof. My proof is complete. □

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Lemma 3.2 (Something)**.** *For every prime p, there is a prime $p' > p$. In particular, there are infinitely many primes.*

Remark 1. Actually, this is a remark.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Corollary 3.2.1 (Anything)**.** *This is my corollary.*

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

A restatement of the theorem is as follows.

Corollary 3.2.1 (Anything)**.** *This is my corollary.*

Definition 3.1 (Apple)**.** An apple is a fruit.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Theorem 3.3 (Keyed theorem)**.** *This is a key-val theorem.*

Theorem 3.3 (continuing from p. 2)**.** *And it's spread out.*

3.1 Direct maximisation

3.2 EM algorithm

3.2.1 EM Algorithm

Substituting (**??**) into (**??**), we can rewrite the I-prior model in a "random-effects" representation. Using matrix notation, we have

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \mathbf{H}_\lambda \mathbf{w} + \boldsymbol{\epsilon} \\ \mathbf{w} &:= (w_1, \ldots, w_n) \sim \mathrm{N}(\mathbf{0}, \psi \mathbf{I}_n) \\ \boldsymbol{\epsilon} &:= (\epsilon_1, \ldots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \psi^{-1} \mathbf{I}_n) \end{aligned} \tag{3.2}$$

where the intercept $\boldsymbol{\alpha} = \alpha \mathbf{1}_n$ has been separated from the regression function. Here, $\mathbf{1}_n$ is a vector of length n containing all ones, and $\mathbf{H}_\lambda$ is the matrix whose (i, j)th entries are $h_\lambda(x_i, x_j)$, where h_λ is the (scaled) reproducing kernel over the set of covariates. An EM algorithm can be applied by treating the random effects $w_1, \ldots, w_n$ as 'missing' in order to estimate the parameters α, $\boldsymbol{\lambda}$ and ψ. The assumption of normality also makes the EM algorithm particularly appealing as the required joint and conditional distributions are easy to obtain. To start, write $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\alpha}, \mathbf{V}_y)$, where $\mathbf{V}_y = \psi \mathbf{H}_\lambda^2 + \psi^{-1} \mathbf{I}_n$ (the marginal distribution of the responses). Given the random effects $\mathbf{w}$, the distribution of $\mathbf{y}|\mathbf{w}$ is also multivariate normal with mean $\boldsymbol{\alpha} + \mathbf{H}_\lambda$ and covariance matrix $\psi^{-1} \mathbf{I}_n$.

The covariance between $\mathbf{y}$ and $\mathbf{w}$ is

$$\begin{aligned} \mathrm{Cov}[\mathbf{y}, \mathbf{w}] &= \mathrm{Cov}[\boldsymbol{\alpha} + \mathbf{H}_\lambda \mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] \\ &= \cancelto{0}{\mathrm{Cov}[\boldsymbol{\alpha}, \mathbf{w}]} + \mathbf{H}_\lambda \cancelto{\psi \mathbf{I}_n}{\mathrm{Cov}[\mathbf{w}, \mathbf{w}]} + \cancelto{0}{\mathrm{Cov}[\boldsymbol{\epsilon}, \mathbf{w}]} \\ &= \psi \mathbf{H}_\lambda := \mathrm{Cov}[\mathbf{w}, \mathbf{y}]. \end{aligned}$$

Thus, the joint distribution of $(\mathbf{y}, \mathbf{w})$ is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{w} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_y & \psi \mathbf{H}_\lambda \\ \psi \mathbf{H}_\lambda & \psi \mathbf{I}_n \end{pmatrix} \right).$$

Using standard results of multivariate normal distributions (see, for example, **krzanowski2000principle**

the conditional distribution of $\mathbf{w}$ given $\mathbf{y}$ is normal with mean and variance

$$\begin{aligned}
\mathrm{E}[\mathbf{w}|\mathbf{y}] &= \mathrm{E}[\mathbf{w}] + \mathrm{Cov}[\mathbf{w},\mathbf{y}]\ \mathrm{Var}[\mathbf{y}]^{-1}\,(\mathbf{y}-\mathrm{E}[\mathbf{y}]) \\
&= \psi\mathbf{H}_\lambda\mathbf{V}_y^{-1}(\mathbf{y}-\boldsymbol{\alpha}) =: \tilde{\mathbf{w}} \\
\mathrm{Var}[\mathbf{w}|\mathbf{y}] &= \mathrm{Var}[\mathbf{w}] + \mathrm{Cov}[\mathbf{w},\mathbf{y}]\ \mathrm{Var}[\mathbf{y}]^{-1}\ \mathrm{Cov}[\mathbf{y},\mathbf{w}] \\
&= \psi\mathbf{I}_n - \psi^2\mathbf{H}_\lambda\mathbf{V}_y^{-1}\mathbf{H}_\lambda \\
&= \mathbf{V}_y^{-1} =: \tilde{\mathbf{V}}_w
\end{aligned} \tag{3.3}$$

where the last equality in the derivation of the conditional variance is obtained using the Woodbury matrix identity. Also, write the second posterior moment of the random effects as

$$\begin{aligned}
\mathrm{E}[\mathbf{w}\mathbf{w}^\top|\mathbf{y}] &= \mathrm{Var}[\mathbf{w}|\mathbf{y}] + \mathrm{E}[\mathbf{w}|\mathbf{y}]\,\mathrm{E}[\mathbf{w}|\mathbf{y}]^\top \\
&= \tilde{\mathbf{V}}_w + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top =: \tilde{\mathbf{W}}.
\end{aligned} \tag{3.4}$$

From the marginal distribution of the responses, we notice that the mean and variance parameters are separable (i.e., they are not dependent on each other). It is straightforward to see that the maximum likelihood estimate for α is $\hat{\alpha} = \bar{y} = \sum_{i=1}^n y_i/n$. We can use this fact and treat the intercept parameter as known.

With g denoting the relevant density functions, the complete data log-likelihood of $\boldsymbol{\lambda}$ and ψ is given by

$$\begin{aligned}
l(\boldsymbol{\lambda},\psi|\mathbf{y},\mathbf{w}) &= \log g(\mathbf{y},\mathbf{w};\hat{\alpha},\boldsymbol{\lambda},\psi) \\
&= \log g(\mathbf{y};\hat{\alpha},\boldsymbol{\lambda},\psi) + \log g(\mathbf{w};\psi) \\
&= -n\log 2\pi - \cancel{\frac{1}{2}\log|\psi^{-1}\mathbf{I}_n|} - \frac{\psi}{2}\|\mathbf{y}-\hat{\boldsymbol{\alpha}}-\mathbf{H}_\lambda\mathbf{w}\|^2 - \cancel{\frac{1}{2}\log|\psi\mathbf{I}_n|} - \frac{\psi}{2}\mathbf{w}^\top\mathbf{w} \\
&= -n\log 2\pi - \frac{\psi}{2}\|\mathbf{y}-\hat{\boldsymbol{\alpha}}\|^2 - \frac{\psi}{2}\mathbf{w}^\top\mathbf{H}_\lambda^2\mathbf{w} + \psi(\mathbf{y}-\hat{\boldsymbol{\alpha}})^\top\mathbf{H}_\lambda\mathbf{w} - \frac{1}{2\psi}\mathbf{w}^\top\mathbf{w} \\
&= -n\log 2\pi - \frac{\psi}{2}\|\mathbf{y}-\hat{\boldsymbol{\alpha}}\|^2 - \frac{1}{2}\mathbf{w}^\top(\underbrace{\psi\mathbf{H}_\lambda^2 + \psi^{-1}\mathbf{I}_n}_{\mathbf{V}_y})\mathbf{w} + \psi(\mathbf{y}-\hat{\boldsymbol{\alpha}})^\top\mathbf{H}_\lambda\mathbf{w} \\
&= -n\log 2\pi - \frac{\psi}{2}\|\mathbf{y}-\hat{\boldsymbol{\alpha}}\|^2 - \frac{1}{2}\,\mathrm{tr}\left(\mathbf{V}_y\mathbf{w}\mathbf{w}^\top\right) + \psi(\mathbf{y}-\hat{\boldsymbol{\alpha}})^\top\mathbf{H}_\lambda\mathbf{w}.
\end{aligned}$$

The EM algorithm at iteration $t \in \{0,1,\dots\}$ entails taking the expectation of the above complete data log-likelihood under $\mathbf{w}$ (the E-step, conditional on the responses $\mathbf{y}$ and some parameter values $(\boldsymbol{\lambda}^{(t)},\psi^{(t)})$. Making use of the results in (3.3) and (3.4), we

denote the tth iteration E-step by the function

$$\begin{aligned} Q(\boldsymbol{\lambda}, \psi) &= \mathrm{E}_{\mathbf{w}}\Big[l(\boldsymbol{\lambda}, \psi | \mathbf{y}, \mathbf{w}) \,\Big|\, \mathbf{y}, \boldsymbol{\lambda}^{(t)}, \psi^{(t)} \Big] \\ &= -n \log 2\pi - \frac{\psi}{2} \|\mathbf{y} - \hat{\boldsymbol{\alpha}}\|^2 - \frac{1}{2} \operatorname{tr}\left(\mathbf{V}_y \tilde{\mathbf{W}}^{(t)} \right) + \psi (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \mathbf{H}_\lambda \tilde{\mathbf{w}}^{(t)}. \end{aligned}$$

The M-step entails maximizing this Q function with respect to the parameters, which then boils down to solving the system of differential equations

$$\frac{\partial Q(\boldsymbol{\lambda}, \psi)}{\partial \lambda_k} = -\frac{1}{2} \operatorname{tr}\left(\frac{\partial \mathbf{V}_y}{\partial \lambda_k} \tilde{\mathbf{W}}^{(t)} \right) + \psi (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \frac{\partial \mathbf{H}_\lambda}{\partial \lambda_k} \tilde{\mathbf{w}}^{(t)} \tag{3.5}$$

$$\frac{\partial Q(\boldsymbol{\lambda}, \psi)}{\partial \psi} = -\frac{1}{2} \|\mathbf{y} - \hat{\boldsymbol{\alpha}}\|^2 - \operatorname{tr}\left(\frac{\partial \mathbf{V}_y}{\partial \psi} \tilde{\mathbf{W}}^{(t)} \right) + (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \mathbf{H}_\lambda \tilde{\mathbf{w}}^{(t)} \tag{3.6}$$

equated to zero for $k = 1, \ldots, p$. The algorithm is made simpler by first conditioning on a value of ψ to obtain updated values for $\boldsymbol{\lambda}$, and then conditioning on these $\boldsymbol{\lambda}$ values to obtain an update for ψ. Given some starting values $(\boldsymbol{\lambda}^{(0)}, \psi^{(0)})$, the E-step and the M-step are iterated until convergence is obtained. A practical stopping criterion would be when there is no longer a sizeable increase in the marginal log-likelihood value, i.e., iterate until

$$\log g(\mathbf{y}; \boldsymbol{\lambda}^{(t+1)}, \psi^{(t+1)}) - \log g(\mathbf{y}; \boldsymbol{\lambda}^{(t)}, \psi^{(t)}) < \delta.$$

for some small value $\delta \in \mathbb{R}$.

For models with iid errors, such as the ones we are considering in this paper, the solution for ψ in (3.6) has the closed-form expression

$$\psi^{(t+1)} = \frac{\operatorname{tr}(\tilde{\mathbf{W}}^{(t)})}{\|\mathbf{y} - \hat{\boldsymbol{\alpha}}\|^2 + \operatorname{tr}(\mathbf{H}_\lambda^2 \tilde{\mathbf{W}}) - 2(\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \mathbf{H}_\lambda \tilde{\mathbf{w}}}.$$

This is generally not the case for the scale parameters $\boldsymbol{\lambda}$, but even so, the system of equations can still be solved using numerical methods. In the next section, we describe cases when closed-form solutions exists for $\boldsymbol{\lambda}$, and how to derive them.

3.2.2 Estimation of the scale parameters

As long as (A) there are no covariates involving square, cubic or any other higher order terms; and (B) the maximum order of interactions between all covariates is two, then the solution to the M-step in (3.5) involving $\boldsymbol{\lambda}$ can be found to be in closed-form. This includes models with either a single or multiple scale parameter(s) with two-way interactions between some or all of the terms. For any other models such as ones involving squared terms and three-way interactions, the M-step can still be solved using numerical methods such as a downhill simplex method. We proceed under the assumptions (A) and (B).

Assume further that there are p covariates, and each of the p kernel matrices $\mathbf{H}_1, \ldots, \mathbf{H}_p$ for the covariates are calculated (depending on whether the data is continuous or nominal). If two-way interactions are present between any $k, j \in \{1, \ldots, p\}$, then these are also calculated as $\mathbf{H}_{kj} = \mathbf{H}_k \circ \mathbf{H}_j$ (the Hadamard product).

Let the number of unique scale parameters be p, i.e., one for each covariate. While the number of scale parameters could actually be less than p, implying that some of the covariates share a scale parameter. This can be thought of as a multi-dimensional covariate. In any case, for a group of such covariates, the kernel matrix is simply the sum of each of the kernel matrices, and all we have to do is re-index everything based on the number of kernel matrices there are, and we are back to p (the number of kernel matrices).

In general, the scaled kernel matrix looks like

$$\mathbf{H}_\lambda = \sum_{k=1}^{p} \lambda_k \mathbf{H}_k + \sum_{k,j \in \mathcal{M}} \lambda_k \lambda_j \mathbf{H}_{kj}$$

where the set $\mathcal{M}$ is the index of all two way interaction terms between the p covariates, i.e., $\mathcal{M} = \{(k, j) : k \text{ interacts with } j, \text{ and } k < j, \ \forall k, j = 1, \ldots, p\}$. Let the number of two-way interactions be $m = |\mathcal{M}|$. The total number of scale parameters is equal to $q = p + m$ when there are non-parsimonious interactions present, otherwise it is $q = p$. The non-parsimonious method of interactions assigns a new scale parameter for each of the Hadamard products of interacting kernel matrices[1]. In comparison, the parsimonious method multiplies the corresponding scale parameters together.

For a particular λ_k, $k = 1, \ldots, q$, we partition the sum of the kernel matrix into parts which involve λ_k and parts which do not:

$$\begin{aligned} \mathbf{H}_\lambda &= \overbrace{\lambda_k \mathbf{H}_k + \lambda_k \sum_{j \in \mathcal{M}} \lambda_j \mathbf{H}_{kj}}^{\lambda_k \text{ is here}} + \overbrace{\sum_{\substack{j=1 \\ j \neq k}}^{p} \lambda_j \mathbf{H}_j + \sum_{\substack{k', j \in \mathcal{M} \\ k' \neq k}} \lambda_{k'} \lambda_j \mathbf{H}_{k'j}}^{\text{no } \lambda_k \text{ here}} \\ &= \lambda_k \mathbf{P}_k + \mathbf{R}_k + \mathbf{U}_k. \end{aligned} \tag{3.7}$$

$\mathbf{P}_k$ is the kernel matrix $\mathbf{H}_k$ plus the sum-product of the interaction kernel matrices with the scale parameters relating to covariate k, i.e., $\sum_j \lambda_j \mathbf{H}_{kj}$. $\mathbf{R}_k$ is the sum-product of the kernel matrices and scale parameters excluding $\lambda_k \mathbf{H}_k$. $\mathbf{U}_k$ is the sum of the interaction cross-product terms excluding those relating to covariate k. Thus, the squared

[1] The non-parsimonious method actually re-indexes both the kernel matrices and scale parameters from $\{1, \ldots, p\}$ to $\{1, \ldots, q\}$ where the Hadamard products are treated like "regular" kernel matrices, and the method proceeds as if there are no interactions present.

kernel matrix is

$$\mathbf{H}_\lambda^2 = \lambda_k^2 \mathbf{P}_k^2 + \lambda_k \left(\mathbf{P}_k \mathbf{R}_k + (\mathbf{P}_k \mathbf{R}_k)^\top + \mathbf{P}_k \mathbf{U}_k + (\mathbf{P}_k \mathbf{U}_k)^\top \right) + \mathbf{R}_k^2 + \mathbf{U}_k^2 + \mathbf{R}_k \mathbf{U}_k + (\mathbf{R}_k \mathbf{U}_k)^\top. \tag{3.8}$$

The M-step from (3.5) for each of the λ_k now reduces to solving

$$\frac{\partial Q(\boldsymbol{\lambda}, \psi)}{\partial \lambda_k} = -\frac{\psi}{2} \operatorname{tr}\left[\left(2\lambda_k \mathbf{P}_k^2 + \mathbf{S}_k \right) \tilde{\mathbf{W}}^{(t)} \right] + \psi (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \mathbf{P}_k \tilde{\mathbf{w}}^{(t)}$$

equal to zero, where we have defined $\mathbf{S}_k = \mathbf{P}_k \mathbf{R}_k + (\mathbf{P}_k \mathbf{R}_k)^\top + \mathbf{P}_k \mathbf{U}_k + (\mathbf{P}_k \mathbf{U}_k)^\top$. This yields the solution

$$\lambda_k^{(t+1)} = \frac{(\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top \mathbf{P}_k \tilde{\mathbf{w}}^{(t)} - \frac{1}{2} \operatorname{tr}\left(\mathbf{S}_k \tilde{\mathbf{W}}^{(t)} \right)}{\operatorname{tr}\left(\mathbf{P}_k^2 \tilde{\mathbf{W}}^{(t)} \right)}$$

for each $k = 1, \ldots, p$.

For most cases, $\mathbf{P}_k$ and $\mathbf{S}_k$ only depend on the kernel matrices and not on the scale parameters, so can be calculated once and stored for efficiency. Further, $\mathbf{U}_k$ equals zero for most cases except in the parsimonious multiple scale parameter case thus simplifying calculations. In fact, we can avoid the expensive matrix multiplications involved in evaluating $\mathbf{P}_k$, its square, and $\mathbf{S}_k$, by calculating once and storing $\mathbf{H}_1, \ldots, \mathbf{H}_p$, its squares and all possible two-way matrix multiplications of these kernel matrices, as the relevant calculation of the M-step merely involves a linear combination of these matrices. These operations are conducted by the `kernL()` function.

3.2.3 Calculation of inverse variance and the likelihood

Sometimes, the calculation of the inverse of $\mathbf{V}_y = \psi \mathbf{H}_\lambda + \psi^{-1} \mathbf{I}_n$ can become problematic, especially when ψ gets extremely large (or extremely small). Notice that $\mathbf{V}_y$ is of the form $\mathbf{A} + s\mathbf{I}_n$, where $\mathbf{A}$ is a symmetric and positive-definite matrix and s is a constant. An eigendecomposition of $\mathbf{A}$ yields $\mathbf{A} = \mathbf{V}\mathbf{U}\mathbf{V}^\top$, where $\mathbf{V}$ is the $n \times n$ matrix whose jth column is the normalised eigenvector $\mathbf{v}_j$ of $\mathbf{A}$ such that $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_n$, and $\mathbf{U}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e., $\mathbf{U}_{ii} = u_i$, for $i = 1, \ldots, n$. Consider then the linear equation

$$\begin{aligned} (\mathbf{A} + s\mathbf{I}_n)\mathbf{a} &= (\mathbf{V}\mathbf{U}\mathbf{V}^\top + s\mathbf{V}\mathbf{V}^\top)\mathbf{a} \\ &= \mathbf{V} \operatorname{diag}(u_1 + s, \ldots, u_n + s)\, \mathbf{V}^\top \mathbf{a} =: \mathbf{b}. \end{aligned}$$

Solving for $\mathbf{a}$ yields

$$\mathbf{a} = \mathbf{V} \operatorname{diag}\left(\frac{1}{u_1 + s}, \ldots, \frac{1}{u_n + s} \right) \mathbf{V}^\top \mathbf{b}. \tag{3.9}$$

With $\mathbf{A} = \psi\mathbf{H}_\lambda^2$ and $s = 1/\psi$, this is a much more stable way of computing $\mathbf{a} = \mathbf{V}_y^{-1}\mathbf{b}$, and is also useful for finding the inverse $\mathbf{a} = \mathbf{V}_y^{-1}$ by setting $\mathbf{b} = \mathbf{I}_n$.

This eigendecomposition is also used for a stable calculation of the log-likelihood. Firstly, the eigenvalues of $\mathbf{A} + s\mathbf{I}_n$ are simply $u_1+s, \ldots, u_n+s$, and the log-determinant can be calculated as

$$\log|\mathbf{A} + s\mathbf{I}_n| = \sum_{i=1}^{n} \log(u_i + s).$$

Furthermore, the matrix $\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\alpha})$ is given by $\mathbf{a}$ in equation (3.9) with $\mathbf{b} = (\mathbf{y} - \boldsymbol{\alpha})$. Thus, the marginal log-likelihood of the responses is given as

$$l(\alpha, \boldsymbol{\lambda}, \psi|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(u_i + s) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})^\top \mathbf{a}.$$

3.2.4 Calculation of standard errors

Consider again the distribution $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\alpha}, \mathbf{V}_y)$, where the mean and covariance matrix depends on different sets of parameters - namely α for the mean, and $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_p, \psi)$ for the covariance matrix. The Fisher information matrix $\mathcal{I}[\alpha, \boldsymbol{\theta}]$ then has the form

$$\mathcal{I}[\alpha, \boldsymbol{\theta}] = \mathrm{diag}\big(\mathcal{I}[\alpha], \mathcal{I}[\boldsymbol{\theta}]\big),$$

where

$$I[\alpha] = \frac{\partial \boldsymbol{\alpha}^\top}{\partial \alpha}\mathbf{V}_y^{-1}\frac{\partial \boldsymbol{\alpha}}{\partial \alpha} = \mathbf{V}_y^{-1} \circ \mathbf{J}_n$$

and

$$I[\boldsymbol{\theta}]_{ij} = \frac{1}{2}\,\mathrm{tr}\left(\mathbf{V}_y^{-1}\frac{\partial \mathbf{V}_y}{\partial \theta_i}\mathbf{V}_y^{-1}\frac{\partial \mathbf{V}_y}{\partial \theta_j}\right).$$

Recall that $\mathbf{V}_y = \psi\mathbf{H}_\lambda^2 + \psi^{-1}\mathbf{I}_n$ and that from (3.8), the derivative of the squared scaled kernel matrix has the form $\partial\mathbf{H}_\lambda^2/\partial\lambda_k = 2\lambda_k\mathbf{P}_k^2 + \mathbf{S}_k$. Therefore, the partial derivative of $\mathbf{V}_y$ with respect to λ_k for $k = 1, \ldots, p$ is

$$\frac{\partial \mathbf{V}_y}{\partial \lambda_k} = \psi\left(2\lambda_k\mathbf{P}_k^2 + \mathbf{S}_k\right),$$

while the partial derivative of $\mathbf{V}_y$ with respect to ψ is

$$\begin{aligned}\frac{\partial \mathbf{V}_y}{\partial \psi} &= \mathbf{H}_\lambda^2 - \frac{1}{\psi^2}\mathbf{I}_n \\ &= \frac{1}{\psi}\left(\psi\mathbf{H}_\lambda^2 + \frac{1}{\psi}\mathbf{I}_n\right) - \frac{2}{\psi^2}\mathbf{I}_n \\ &= \frac{1}{\psi}\mathbf{V}_y - \frac{2}{\psi^2}\mathbf{I}_n.\end{aligned}$$

The Fisher information matrix can be obtained fairly inexpensively as the two matrices $\mathbf{P}_k^2$ and $\mathbf{S}_k$ have already been calculated through the EM algorithm (if the $\boldsymbol{\lambda}$ are in closed-form). Otherwise, there exist R functions to compute the Hessian of the (negative) log-likelihood numerically. The standard error for $\eta \in \{\alpha, \lambda_1, \ldots, \lambda_p, \psi\} =: \mathcal{E}$ is then given by

$$\text{s.e.}(\eta) = \sqrt{\left(I[\alpha, \boldsymbol{\theta}]^{-1}\right)_{kk}},$$

where k is the index of the parameter in the set $\mathcal{E}$. The **iprior** package employs Wald tests of significance based off of these standard errors for the intercept and scale parameters in the `summary` of an `ipriorMod` object.

3.3 The EM algorithm pseudocode

Something I just realised...

Already have the eigendecomposition $\mathbf{H}_\lambda = \mathbf{V}\,\text{diag}(u_1, \ldots, u_n)\mathbf{V}^\top$ with $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$. Also, $\tilde{\mathbf{W}} = \mathbf{V}_y^{-1} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top$, and $\mathbf{V}_y = \psi\mathbf{H}_\lambda^2 + \psi^{-1}\mathbf{I}_n$. Therefore $\mathbf{V}_y^{-1} = \mathbf{V}\,\text{diag}\left(\frac{1}{\psi u_i^2 + 1/\psi}\right)\mathbf{V}^\top$. Then,

$$\begin{aligned}\text{tr}(\mathbf{H}_\lambda^2\tilde{\mathbf{W}}) &= \text{tr}\left(\mathbf{H}_\lambda^2(\mathbf{V}_y^{-1} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top)\right) \\ &= \text{tr}(\mathbf{H}_\lambda^2\mathbf{V}_y^{-1}) + \text{tr}(\mathbf{H}_\lambda^2\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top) \\ &= \text{tr}\left(\mathbf{V}\,\text{diag}(u_i)\cancel{\mathbf{V}^\top\mathbf{V}}\,\text{diag}\left(\frac{1}{\psi u_i^2 + 1/\psi}\right)\mathbf{V}^\top\right) + \tilde{\mathbf{w}}^\top\mathbf{H}_\lambda\mathbf{H}_\lambda\tilde{\mathbf{w}} \\ &= \text{tr}\left(\cancel{\mathbf{V}^\top\mathbf{V}}\,\text{diag}\left(\frac{u_i}{\psi u_i^2 + 1/\psi}\right)\right) + \tilde{\mathbf{w}}^\top\mathbf{H}_\lambda\mathbf{H}_\lambda\tilde{\mathbf{w}} \\ &= \sum_{i=1}^n \frac{u_i}{\psi u_i^2 + 1/\psi} + \tilde{\mathbf{w}}^\top\mathbf{H}_\lambda\mathbf{H}_\lambda\tilde{\mathbf{w}}\end{aligned}$$

This way, no need to square any matrices.

Here is the pseudocode for the EM algorithm to estimate the I-prior model. An estimate of the percentage run time for some of the heavier computations are also given.

Algorithm 1 EM algorithm for the I-prior model

procedure INITIALISE (PART OF THE KERNEL LOADER)
  Choose suitable $\boldsymbol{\lambda}^{(0)} = (\lambda_1^{(0)}, \ldots, \lambda_l^{(0)})$ and $\psi^{(0)}$
  $t \leftarrow 0$
  $\hat{\alpha} \leftarrow \sum_{i=1}^{n} y_i/n$ ▷ The MLE for $\hat{\alpha}$
  $p \leftarrow$ no. of covariates
  $m \leftarrow$ no. of interactions
  $q \leftarrow$ no. of expanded scale parameters/kernel matrices $(p + m)$
  for $k = 1, \ldots, p$ **do**
    Calculate $\mathbf{H}_k, \mathbf{H}_k^2$, and any Hadamard products (interactions).
    Calculate $\mathbf{P}_k, \mathbf{P}_k^2, \mathbf{S}_k$ using kernel matrices $\mathbf{H}_k$ and $\boldsymbol{\lambda}^{(0)}$. ▷ time: 3.4%
  end for
  Index all the relevant kernel matrices from 1 to q.
end procedure

procedure BLOCK A UPDATE (Iteration 0)
  Expand $\boldsymbol{\lambda}_{1:l}^{(t)} \to \boldsymbol{\lambda}_{1:q}^{(t)}$ depending on interactions and higher order terms
  $\mathbf{H}_\lambda \leftarrow \sum_{k=1}^{q} \boldsymbol{\lambda}_k^{(t)} \mathbf{H}_k$
  procedure EIGENDECOMPOSITION ▷ time: 51.6%
    $(\text{diag}(u_1, \ldots, u_n), \mathbf{V}) \leftarrow \text{eigen}(\mathbf{H}_\lambda)$
  end procedure
  $s \leftarrow 1/\psi^{(t)}$
end procedure

function LINEAR SOLVER AND INVERSE (input $\mathbf{b}$)
  $\mathbf{a} \leftarrow \mathbf{V}\,\text{diag}\left[\frac{s}{u_1^2+s^2}, \ldots, \frac{s}{u_n^2+s^2}\right]\mathbf{V}^\top\mathbf{b}$
end function

procedure LOG-LIKELIHOOD UPDATE
  call LINEAR SOLVER AND INVERSE (input $\mathbf{b} = \mathbf{y} - \hat{\boldsymbol{\alpha}}$)
  $l \leftarrow -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(u_i^2/s + s) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})^\top\mathbf{a}$
end procedure

procedure BLOCK B UPDATE (t)
    Update $\mathbf{P}_k$ and $\mathbf{S}_k$. ▷ time: 15.6%
end procedure

while $l_{new} - l_{old} > \delta$ **or** $t < t_{max}$ **do** ▷ The EM iterations
    procedure BLOCK C UPDATE (Iteration t)
        call LINEAR SOLVER AND INVERSE (input $\mathbf{b} = \mathbf{I}_n$) ▷ time: 10.4%
        $\mathbf{V}_y^{-1} \leftarrow \mathbf{a}$
        $\tilde{\mathbf{w}} \leftarrow \psi^{(t)}\mathbf{H}_\lambda\mathbf{V}_y^{-1}(\mathbf{y} - \hat{\boldsymbol{\alpha}})$
        $\tilde{\mathbf{W}} \leftarrow \mathbf{V}_y^{-1} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top$
    end procedure

    procedure UPDATE FOR $\boldsymbol{\lambda}$: CLOSED-FORM EM
        for $k = 1, \ldots, l$ **do**
            $T_1 \leftarrow \operatorname{tr}\left(\mathbf{P}_k^2\tilde{\mathbf{W}}\right)$
            $T_2 \leftarrow (\mathbf{y} - \hat{\boldsymbol{\alpha}})\mathbf{P}_k\tilde{\mathbf{w}} - \frac{1}{2}\operatorname{tr}\left(\mathbf{S}_k\tilde{\mathbf{W}}\right)$ ▷ time: 8.5%
            $\lambda_k^{(t+1)} \leftarrow T_2/T_1$
        end for
    end procedure
    Note: If higher order terms and/or three-way (or more) interactions are present, then a numerical method is used.

    procedure UPDATE FOR ψ ▷ time: 5.8%
        $T_3 \leftarrow (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top(\mathbf{y} - \hat{\boldsymbol{\alpha}}) + \operatorname{tr}(\mathbf{H}_\lambda^2\tilde{\mathbf{W}}) - 2(\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top\mathbf{H}_\lambda\tilde{\mathbf{w}}$
        $\psi^{(t+1)} \leftarrow \sqrt{\operatorname{tr}(\tilde{\mathbf{W}})/T_3}$
    end procedure

    call BLOCK A UPDATE (Iteration $t + 1$)
    call LOG-LIKELIHOOD UPDATE
    $t \leftarrow t + 1$
end while

$(\hat{\boldsymbol{\lambda}}, \hat{\psi}) \leftarrow (\boldsymbol{\lambda}^{(t)}, \psi^{(t)})$ ▷ The maximum likelihood estimates

3.4 Markov chain Monte Carlo methods

3.5 Low-rank matrix approximation (Nyström method)

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Symbols

$\mathrm{N}_p(\mu, \Sigma)$	p-dimensional multivariate normal distribution with mean vector μ and covariance Σ.
$\sim$	Is distributed as.
$\otimes$	The tensor product.

Index