I define the following variables:

X — The matrix $m \times n$ of predictors

$\mathbf{y}$ — The response m-vector

B — The current $m \times q$ basis matrix

W — The weight matrix ($W = \sqrt{\Omega^{-1}}$). The usual case is that W is diagonal (that's what py-earth currently supports). W is not necessarily symmetric, although Ω always is. For example, W could be a Cholesky factor of Ω^{-1} and would therefore be triangular.

Q — An orthonormalized B. $Q = BT$ and $Q^T W^T W Q = I_n$ and $\operatorname{span}(Q) = \operatorname{span}(B)$

T — The upper triangular $q \times q$ matrix that orthonormalizes Q. T is the cholesky factor of $B^T W^T W B$.

D — Composite matrix $[B, \mathbf{y}]$

M — Orthonormalized D. $M = DS$ and $M^T W^T W M = I_n$ and $\operatorname{span}(M) = \operatorname{span}(D)$ and $M = [Q, \mathbf{z}]$

S — The upper triangular $(q+1) \times (q+1)$ matrix that orthonormalizes D. S is the cholesky factor of $D^T W^T W D$.

$\mathbf{z}$ — The orthonormalized $\mathbf{y}$. That is, $B^T \mathbf{z} = Q^T \mathbf{z} = \mathbf{0}$ and $\mathbf{y}^T \mathbf{y} = 1$ and $\operatorname{span}([B, \mathbf{y}]) = \operatorname{span}([Q, \mathbf{z}])$.

$\mathbf{b}$ — The candidate new basis vector. $\mathbf{b}_i = \max(\mathbf{p}_i X_{i,c} - \phi, 0) = \max(\mathbf{p}_i \mathbf{x}_i - \phi, 0)$ for some knot candidate, ϕ, and variable candidate, c, and parent candidate vector, $\mathbf{p} = B_{:,d}$ for parent candidate d.

$\mathbf{x}$ — The candidate variable vector. $\mathbf{x} = X_{:,c}$.

c — The candidate variable index. See $\mathbf{x}$.

$\mathbf{p}$ — The candidate parent vector. $\mathbf{p} = B_{:,d}$.

d — The candidate parent index. See $\mathbf{p}$.

$\mathbf{c}$ — The weighted candidate new basis vector, $\mathbf{c} = W\mathbf{b}$.

I'll use the following conventions. All non-bold capital letters are matrices. All bold lowercase letters are vectors. All lowercase non-bold letters are scalars. If I make a rank 3 tensor, it will be uppercase and bold. I'm sure I won't need to go past rank 3. I'll represent a row of a matrix, Φ, by $\Phi_{i,:}$, a column by $\Phi_{:,j}$ and a sub-matrix by $\Phi_{a:b,c:d}$ and similarly for vectors. An element of Φ is $\Phi_{i,j}$. If there are no ranges involved, I might drop the comma and just write Φ_{ij}. If I'm using actual numbers, or it's in any way ambiguous, I'll keep the comma.

I'll use i to index rows and j to index columns. I'll use k and h for rows and columns, respectively, if I need additional indices.

Let $V = [Q, \mathbf{c}, \mathbf{z}]^T [Q, \mathbf{c}, \mathbf{z}]$ and let C be its upper triangular cholesky factor such that $V = C^T C$. Then V and C have the following special structures:

$$V = \begin{bmatrix} I_q & \gamma & \mathbf{0} \\ \gamma^T & \beta & \alpha \\ \mathbf{0} & \alpha & 1 \end{bmatrix} \tag{1}$$

$$C = \begin{bmatrix} I_q & \delta & \mathbf{0} \\ \mathbf{0} & \epsilon & \zeta \\ \mathbf{0} & 0 & \eta \end{bmatrix} \tag{2}$$

The following identities hold:

$$\zeta^2 + \eta^2 = 1 \tag{3}$$
$$\zeta\epsilon = \alpha \tag{4}$$
$$\gamma = \delta \tag{5}$$
$$\beta = \delta^T\delta + \epsilon^2 \tag{6}$$
$$= \delta^2 + \epsilon^2 \tag{7}$$
$$= \gamma^2 + \epsilon^2 \tag{8}$$

and their inverses are:

$$\eta = \sqrt{1 - \zeta^2} \tag{9}$$
$$\zeta = \alpha/\epsilon \tag{10}$$
$$\epsilon = \sqrt{\beta - \delta^2} \tag{11}$$
$$= \sqrt{\beta - \gamma^2} \tag{12}$$

The objective here is to minimize η, which is the root mean squared error of the solution to the least squares problem

$$\eta = \min_{\psi \in \mathbb{R}^{q+1}} \sqrt{\left([Q, \mathbf{c}]\,\psi - \mathbf{z}\right)^2} \tag{13}$$

which is equivalent to the objective of the weighted least squares problem we want to solve. Let's say there are two candidate knots, ϕ and $\tilde{\phi}$, $\tilde{\phi} < \phi$. All quantities discussed so far relate to ϕ. I want to compute the corresponding $\tilde{}$ quantities, associated with $\tilde{\phi}$, from the original quantities as quickly as possible. There is a fast update rule for $\mathbf{b}$, which is

$$\tilde{\mathbf{b}}_i - \mathbf{b}_i = \begin{cases} 0, & \mathbf{x}_i \leq \tilde{\phi} \\ \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right), & \tilde{\phi} < \mathbf{x}_i < \phi \\ \mathbf{p}_i \left(\phi - \tilde{\phi}\right) & \mathbf{x}_i \geq \phi \end{cases} \tag{14}$$

Let's get the update formulas for α, β, γ.

$$\alpha = \mathbf{b}^T W^T \mathbf{z} \tag{15}$$
$$= \sum_{i=1}^{m} \mathbf{w}_i \mathbf{b}_i \mathbf{z}_i \tag{16}$$
$$\Delta\alpha = \sum_{i=1}^{m} \mathbf{w}_i \mathbf{z}_i \Delta \mathbf{b}_i \tag{17}$$
$$= \sum_{i=1}^{m} \mathbf{w}_i \mathbf{z}_i \begin{cases} 0, & \mathbf{x}_i \leq \tilde{\phi} \\ \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right), & \tilde{\phi} < \mathbf{x}_i < \phi \\ \mathbf{p}_i \left(\phi - \tilde{\phi}\right) & \mathbf{x}_i \geq \phi \end{cases} \tag{18}$$
$$= \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right) + \sum_{i:\mathbf{x}_i \geq \phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \left(\phi - \tilde{\phi}\right) \tag{19}$$
$$= \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right) + \sum_{i:\mathbf{x}_i \geq \phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \left(\phi - \tilde{\phi}\right) \tag{20}$$
$$= \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \mathbf{x}_i - \tilde{\phi} \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i + \left(\phi - \tilde{\phi}\right) \sum_{i:\mathbf{x}_i \geq \phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \tag{21}$$
$$= \sigma - \tilde{\phi}\tau + \left(\phi - \tilde{\phi}\right) \upsilon \tag{22}$$

where

$$\sigma = \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \mathbf{x}_i \tag{23}$$
$$\tau = \sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \tag{24}$$
$$\upsilon = \sum_{i:\mathbf{x}_i \geq \phi} \mathbf{w}_i \mathbf{z}_i \mathbf{p}_i \tag{25}$$

For β,

$$\begin{aligned}
\beta &= \mathbf{b}^T W^T W \mathbf{b} && (26)\\
&= \sum_{i=1}^{m} \mathbf{w}_i^2 \mathbf{b}_i^2 && (27)\\
\Delta\beta &= \sum_{i=1}^{m} \mathbf{w}_i^2 \tilde{\mathbf{b}}_i^2 - \sum_{i=1}^{m} \mathbf{w}_i^2 \mathbf{b}_i^2 && (28)\\
&= \sum_{i=1}^{m} \mathbf{w}_i^2 \left(\tilde{\mathbf{b}}_i^2 - \mathbf{b}_i^2\right) && (29)\\
&= \sum_{i=1}^{m} \mathbf{w}_i^2 \left(\tilde{\mathbf{b}}_i + \mathbf{b}_i\right)\left(\tilde{\mathbf{b}}_i - \mathbf{b}_i\right) && (30)\\
&= \sum_{i=1}^{m} \mathbf{w}_i^2 \left(\mathbf{b}_i + \Delta\mathbf{b}_i + \mathbf{b}_i\right)\left(\mathbf{b}_i + \Delta\mathbf{b}_i - \mathbf{b}_i\right) && (31)\\
&= \sum_{i=1}^{m} \mathbf{w}_i^2 \left(2\mathbf{b}_i + \Delta\mathbf{b}_i\right)\Delta\mathbf{b}_i && (32)\\
&= 2\sum_{i=1}^{m} \mathbf{w}_i^2 \mathbf{b}_i \Delta\mathbf{b}_i + \sum_{i=1}^{m} \mathbf{w}_i^2 \left(\Delta\mathbf{b}_i\right)^2 && (33)\\
&= 2\sum_{i=1}^{m} \mathbf{w}_i^2 \mathbf{b}_i \begin{cases} 0, & \mathbf{x}_i \le \tilde{\phi} \\ \mathbf{p}_i\left(\mathbf{x}_i - \tilde{\phi}\right), & \tilde{\phi} < \mathbf{x}_i < \phi \\ \mathbf{p}_i\left(\phi - \tilde{\phi}\right) & \mathbf{x}_i \ge \phi \end{cases} && (34)\\
&\quad + \sum_{i=1}^{m} \mathbf{w}_i^2 \begin{cases} 0, & \mathbf{x}_i \le \tilde{\phi} \\ \mathbf{p}_i^2\left(\mathbf{x}_i - \tilde{\phi}\right)^2, & \tilde{\phi} < \mathbf{x}_i < \phi \\ \mathbf{p}_i^2\left(\phi - \tilde{\phi}\right)^2 & \mathbf{x}_i \ge \phi \end{cases} && (35)\\
&= 2\sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi}^{m} \mathbf{w}_i^2 \mathbf{b}_i \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right) && (36)\\
&\quad +2\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \mathbf{b}_i \mathbf{p}_i \left(\phi - \tilde{\phi}\right) && (37)\\
&\quad +\sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \left(\mathbf{x}_i - \tilde{\phi}\right)^2 && (38)\\
&\quad +\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \left(\phi - \tilde{\phi}\right)^2 && (39)\\
&= 0 + 2\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \left(\mathbf{x}_i - \phi\right)\mathbf{p}_i^2 \left(\phi - \tilde{\phi}\right) && (40)\\
&\quad +\sum_{i:\tilde{\phi}<\mathbf{x}_i<\phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \left(\mathbf{x}_i - \tilde{\phi}\right)^2 && (41)\\
&\quad +\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \left(\phi - \tilde{\phi}\right)^2 && (42)\\
&= 2\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \mathbf{x}_i \phi - 2\sum_{i:\mathbf{x}_i\ge\phi}^{m} \mathbf{w}_i^2 \mathbf{x}_i \mathbf{p}_i^2 \tilde{\phi} && (43)
\end{aligned}$$

where

$$\lambda = \sum_{i:\mathbf{x}_i \geq \phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \mathbf{x}_i \quad (68)$$

$$\mu = \sum_{i:\mathbf{x}_i \geq \phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \quad (69)$$

$$\nu = \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \quad (70)$$

$$\xi = \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \mathbf{x}_i \quad (71)$$

$$\rho = \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} \mathbf{w}_i^2 \mathbf{p}_i^2 \mathbf{x}_i^2 \quad (72)$$

For γ,

$$\begin{aligned}
\gamma &= Q^T W \mathbf{b} && (73)\\
\gamma_j &= Q_{j,:}^T W \mathbf{b} && (74)\\
&= \sum_{i=1}^{m} Q_{ij} \mathbf{w}_i \mathbf{b}_i && (75)\\
&= \sum_{i=1}^{m} Q_{ij} \mathbf{w}_i \begin{cases} 0, & \mathbf{x}_i \leq \tilde{\phi} \\ \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right), & \tilde{\phi} < \mathbf{x}_i < \phi \\ \mathbf{p}_i \left(\phi - \tilde{\phi}\right) & \mathbf{x}_i \geq \phi \end{cases} && (76)\\
&= \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} Q_{ij} \mathbf{w}_i \mathbf{p}_i \left(\mathbf{x}_i - \tilde{\phi}\right) && (77)\\
&\quad + \sum_{i:\mathbf{x}_i \geq \phi}^{m} Q_{ij} \mathbf{w}_i \mathbf{p}_i \left(\phi - \tilde{\phi}\right) && (78)\\
&= \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} Q_{ij} \mathbf{w}_i \mathbf{p}_i \mathbf{x}_i - \tilde{\phi} \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} Q_{ij} \mathbf{w}_i \mathbf{p}_i && (79)\\
&\quad + \left(\phi - \tilde{\phi}\right) \sum_{i:\mathbf{x}_i \geq \phi}^{m} Q_{ij} \mathbf{w}_i \mathbf{p}_i && (80)\\
&= \chi_j - \tilde{\phi} \psi_j + \left(\phi - \tilde{\phi}\right) \kappa_j && (81)
\end{aligned}$$

where

$$\kappa_j = \sum_{i:\mathbf{x}_i \geq \phi}^{m} Q_{ij}\mathbf{w}_i\mathbf{p}_i \tag{82}$$

$$\chi_j = \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} Q_{ij}\mathbf{w}_i\mathbf{p}_i\mathbf{x}_i \tag{83}$$

$$\psi_j = \sum_{i:\tilde{\phi} < \mathbf{x}_i < \phi}^{m} Q_{ij}\mathbf{w}_i\mathbf{p}_i \tag{84}$$

Converting back to problem scale

Suppose you don't orthonormalize $\mathbf{y}$ with B. Then let $V' = [Q, \mathbf{c}, \mathbf{y}]^T [Q, \mathbf{c}, \mathbf{y}]$ and let C be its upper triangular cholesky factor such that $V' = C'^T C'$. Then V' and C' have the following special structures:

$$V' = \begin{bmatrix} I_q & \gamma & \theta \\ \gamma^T & \beta & \alpha' \\ \theta^T & \alpha' & \omega \end{bmatrix} \tag{85}$$

$$C' = \begin{bmatrix} I_q & \delta & \iota \\ \mathbf{0} & \epsilon & \zeta' \\ \mathbf{0} & 0 & \eta' \end{bmatrix} \tag{86}$$

The following identities hold:

$$\iota^2 + \zeta'^2 + \eta'^2 = \omega \tag{87}$$

$$\zeta'\epsilon + \delta^T\iota = \alpha' \tag{88}$$

$$\theta = \iota \tag{89}$$

$$\gamma = \delta \tag{90}$$

$$\beta = \gamma^2 + \epsilon^2 \tag{91}$$

and their inverses are:

$$\eta' = \sqrt{\omega - \zeta'^2 - \iota^2} \tag{92}$$

$$= \sqrt{\omega - \zeta'^2 - \theta^2} \tag{93}$$

$$\zeta' = \frac{\alpha' - \delta^T\theta}{\epsilon} \tag{94}$$

$$= \frac{\alpha' - \gamma^T\theta}{\epsilon} \tag{95}$$

$$\epsilon = \sqrt{\beta - \gamma^2} \tag{96}$$

We can do the same fast updates as before with $W\mathbf{y}$ in place of $\mathbf{z}$ to get α' and proceed from there to solve the problem without orthonormalizing the outcome.

If we take out $\mathbf{c}$ and just do linear regression for some reason (this comes up in the actual code for the forward pass at various steps), we get the simpler system

$$V = [Q, \mathbf{y}]^T [Q, \mathbf{y}] = \begin{bmatrix} I_q & \theta \\ \theta^T & \omega \end{bmatrix} \tag{97}$$

$$C = \begin{bmatrix} I_q & \theta \\ \mathbf{0}_q^T & \sqrt{\omega - \theta^T\theta} \end{bmatrix} \tag{98}$$

which can be confirmed by staring at this equation:

$$\begin{bmatrix} I_q & \mathbf{0}_q \\ \theta^T & \sqrt{\omega - \theta^T\theta} \end{bmatrix} \begin{bmatrix} I_q & \theta \\ \mathbf{0}_q & \sqrt{\omega - \theta^T\theta} \end{bmatrix} = \begin{bmatrix} I_q & \theta \\ \theta^T & \omega \end{bmatrix} \tag{99}$$

Then $\min_{\psi \in \mathbb{R}^q} \sqrt{([Q]\,\psi - \mathbf{z})^2} = \sqrt{\omega - \theta^T\theta}$.

Multiple responses

Suppose we have an $m \times p$ matrix Y instead of vector $\mathbf{y}$. Let W be an $m \times p$ matrix of corresponding weights and $\mathbf{W}_{:,:,j} = \operatorname{diag}(W_{:,j})$. Okay. The math is basically the same, but needs to be repeated for each outcome. That means most tensors increase in rank. Most importantly, Q becomes the $m \times q \times p$ tensor $\mathbf{Q}$, with $\mathbf{Q}_{:,:,j}$ corresponding to $W_{:,j}$. For each knot candidate, the total ζ, which is the sum of the ζ for each outcome, is the quantity to maximize.

Let $\mathbf{V}$ be an $q \times q \times p$ tensor such that

$$\mathbf{V}_{:,:,k} = \begin{bmatrix} I_q & \Gamma_{:,k} & \Theta_{:,k} \\ (\Gamma_{:,k})^T & \beta_k & \alpha_k \\ (\Theta_{:,k})^T & \alpha_k & \omega_k \end{bmatrix} \tag{100}$$

and let

$$\mathbf{C}_{:,:,k} = \begin{bmatrix} I_q & \Delta_{:,k} & \iota_{:,k} \\ \mathbf{0} & \epsilon_k & \zeta_k \\ \mathbf{0} & 0 & \eta_k \end{bmatrix} \tag{101}$$

The same identities hold

$$\iota_{:,k}^2 + \zeta_k^2 + \eta_k^2 = \omega_k \tag{102}$$

$$\zeta_k \epsilon_k + \Delta_{:,k}^T \iota_{:,k} = \alpha_k \tag{103}$$

$$\Theta_{:,k} = \iota_{:,k} \tag{104}$$

$$\boldsymbol{\Gamma}_{:,k} = \Delta_{:,k} \tag{105}$$

$$\beta_k = \Gamma_{:,k}^2 + \epsilon_k^2 \tag{106}$$

with inverses

$$\eta_k = \sqrt{\omega_k - \zeta_k^2 - \iota_{:,k}^2} \tag{107}$$

$$= \sqrt{\omega_k - \zeta_k^2 - \Theta_{:,k}^2} \tag{108}$$

$$\zeta_k = \frac{\alpha_k - \Gamma_{:,k}^T \Theta_{:,k}}{\epsilon_k} \tag{109}$$

$$\epsilon_k = \sqrt{\beta_k - \Gamma_{:,k}^2} \tag{110}$$

The objective is to minimize $\eta^2 = \sum_{k=1}^p \eta_k^2$, which is equivalent to maximizing $\zeta^2 = \sum_{k=1}^p \zeta_k^2$ (because ω and Θ do not depend on the knot value).

$$\zeta^2 = \sum_{k=1}^p \zeta_k^2 \tag{111}$$

$$= \sum_{k=1}^p \frac{\alpha_k - \Gamma_{:,k}^T \Theta_{:,k}}{\epsilon_k} \tag{112}$$

$$= \sum_{k=1}^p \frac{\alpha_k}{\epsilon_k} - \sum_{k=1}^p \frac{\Gamma_{:,k}^T \Theta_{:,k}}{\epsilon_k} \tag{113}$$

$$= \sum_{k=1}^p \frac{\alpha_k}{\beta_k - \Gamma_{:,k}^2} - \sum_{k=1}^p \frac{\Gamma_{:,k}^T \Theta_{:,k}}{\beta_k - \Gamma_{:,k}^2} \tag{114}$$