Notes on Statistical and Machine Learning

Generalized Additive Models

Chapter: *12* **Prepared by:** *Chenxi Zhou*

This note is produced based on

- *Chapter 9, Additive Models, Trees, and Related Methods* in Hastie, Tibshirani, and Friedman (2009), and
- *Chapter 4, Generalizations of the Lasso Penalty* in Hastie, Tibshirani, and Wainwright (2015).

I. Generalized Additive Models

1. **Generalized Additive Model for Gaussian Response:** In the regression setting, the *generalized additive model* has the form

$$\mathbb{E}[Y \mid X_1, X_2, \cdots, X_p] = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p), \tag{1}$$

where $X_1, \cdots, X_p$ are the predictors, Y is the response assumed to follow a Gaussian distribution, and f_j's are unspecified smooth functions.

2. **Generalized Additive Logistic Regression Model:** Suppose that the response variable is binary.

   (a) *Review of classic logistic regression model:* In the classic logistic regression model, we relate the mean of the binary response $\mu(X) = \mathbb{P}(Y = 1 \mid X)$ to predictors using the *logit* link function

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2}$$

   (b) *Generalized Additive Logistic Regression Model:* The *generalized additive logistic regression model* replaces each linear term by a more general functional form

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p), \tag{3}$$

   where each f_j is an unspecified smooth function.

   (c) *Comparison Between (2) and (3):*

      i. The additivity is retained, which makes the interpretation easy;
      ii. We allow $\log\left(\frac{\mu(X)}{1-\mu(X)}\right)$ to depend on each of $X_1, \cdots, X_p$ in a nonparametric fashion, allowing the model to be more flexible.

3. **Generalized Additive Model in General Settings:** We let the conditional mean $\mu(X)$ of a response Y be related to an additive function of the predictors via a *link function* g. The resulting generalized additive model is

$$g(\mu(X)) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

Examples:

- $g(\mu) = \mu$ is the *identity link*, used for linear and additive models for Gaussian response data;
- $g(\mu) = \text{logit}(\mu)$, or $g(\mu) = \text{probit}(\mu)$, the *probit link function*, for modeling binomial probabilities, where the probit function is the inverse Gaussian cumulative distribution function,

$$\text{probit}(\mu) = \Phi^{-1}(\mu). \tag{4}$$

- $g(\mu) = \log(\mu)$ for log-linear or log-additive models for Poisson count data.

4. **Some Extensions on $\{f_j\}_{j=1}^p$:**

   (a) By letting f_j's to be nonlinear in X_j's, the estimated function can then reveal possible nonlinearities in the effect of $X_1, \cdots, X_p$;

   (b) *Not* all of the functions f_j's need to be nonlinear. We can mix in linear and other parametric forms with the nonlinear terms.

   (c) The nonlinear terms are *not* restricted to *main effects* either; we can have nonlinear components in two or more variables, or separate curves in X_j for each level of the factor X_k.

   (d) *Examples:*

   - $g(\mu(X)) = X^\top \boldsymbol{\beta} + \alpha_k + f(Z)$ — a semiparametric model, where
     - X is a vector of predictors to be modeled linearly,
     - α_k the effect for the k-th level of a qualitative predictor, and
     - the effect of predictor Z is modeled nonparametrically;
   - $g(\mu(X)) = f(X) + g_k(Z)$ — k indexes the levels of a qualitative input V, and thus creates an interaction term $g(V, Z) = g_k(Z)$ for the effects of V and Z;
   - $g(\mu(X)) = f(X) + g(Z, W)$, where g is a nonparametric function in two features.

5. **Fitting Generalized Additive Models Under Squared-error Loss:**

   (a) *Model specification:* We assume the response Y and the predictors $X_1, \cdots, X_p$ are related by

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon,$$

   where $\varepsilon \sim \text{Normal}(0, \sigma^2)$ and $\sigma^2 > 0$ is unknown.

(b) *Fitting criterion:* Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be i.i.d data, where $\mathbf{x}_i \in \mathbb{R}^p$. We minimize the penalized residual sum-of-squares (PRSS)

$$\begin{aligned}&\text{PRSS}_{\lambda_1,\cdots,\lambda_p}(\alpha, f_1, \cdots, f_p)\\ &:= \sum_{i=1}^n \left(y_i - \alpha - \sum_{j=1}^p f_j(x_{i,j})\right)^2 + \sum_{j=1}^p \lambda_j \int (f_j''(t_j))^2 \mathrm{d}t_j, \end{aligned} \tag{5}$$

where $\lambda_j \geq 0$ are tuning parameters.

(c) *Characterizing the Solution to* (5)*:* The minimizer to (5) is an *additive cubic spline model*.

i. Each f_j is a cubic spline in the component X_j, with knots at each of the unique values of $x_{i,j}$, for all $i = 1, \cdots, n$. Without further restrictions on the model, the solution is *not* unique.

ii. The constant α is not identifiable, since we can add or subtract any constants to each of the functions f_j, and adjust α accordingly.

(d) *Restrictions on f_j's:* Assume $\sum_{i=1}^n f_j(x_{i,j}) = 0$ for all $j = 1, \cdots, p$, i.e., the functions average zero over data. Then,

i. the minimizer of α in (5) is $\hat{\alpha} = \text{Ave}(y_i)$;

ii. if, in addition, the matrix of input values (whose (i, j)-th entry is $x_{i,j}$) has full column rank, then (5) is a strictly convex criterion and the minimizer is unique.

(e) *Backfitting algorithm:* Under the restrictions mentioned above, with $\hat{\alpha} = \text{Ave}(y_i)$, we apply a cubic smoothing spline $\mathcal{S}_j$ to the targets

$$y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{i,k}), \qquad \text{for all } i = 1, \cdots, n,$$

to obtain a new estimate $\hat{f}_j$. This procedure is done for each predictor in turn, using the current estimates of $\hat{f}_k$, where $k \neq j$. The process is continued until the estimates $\hat{f}_j$'s stabilize, for all $j = 1, 2, \cdots, p$.

The complete algorithm is provided in Algorithm 1.

Algorithm 1 Backfitting Algorithm for Generalized Additive Model

1: Initialize $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^n y_i$, $\hat{f}_j = 0$ for all $j = 1, \cdots, p$;
2: Cycle $j = 1, 2, \cdots, p, 1, 2, \cdots, p, \cdots$,

$$\begin{aligned}\hat{f}_j &\leftarrow \mathcal{S}_j\left[\left\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{i,k})\right\}_{i=1}^n\right],\\ \hat{f}_j &\leftarrow \hat{f}_j - \frac{1}{n}\sum_{i=1}^n \hat{f}_j(x_{i,j}).\end{aligned}$$

until the functions $\hat{f}_j$ change less than a pre-specified threshold.

(f) *Degrees of freedom:* If we consider the operation of smoother $\mathcal{S}_j$ only at the training points, it can be represented by an $n \times n$ operator matrix $\mathbf{S}_j$. Then the degrees of freedom for the j-th term are (approximately) computed as

$$\mathrm{df}_j = \mathrm{trace}(\mathbf{S}_j) - 1. \tag{6}$$

6. Fitting Algorithms in General:

(a) *Criterion:* For the logistic regression model and other generalized additive models, we maximize the *penalized log-likelihood function*.

(b) *Backfitting algorithm:* The backfitting procedure is used in conjunction with a likelihood maximizer. The usual Newton-Raphson routine for maximizing log-likelihoods can be recast as an *iteratively reweighted least squares (IRLS)* algorithm.

(c) *Backfitting algorithm for generalized additive logistic model:*

Algorithm 2 Local Scoring Algorithm for the Additive Logistic Regression Model

1: Compute starting values: $\hat{\alpha} = \log(\bar{y}/(1-\bar{y}))$, where $\bar{y} = \mathrm{Ave}(y_i)$, the sample proportion of ones, and set $\hat{f}_j = 0$ for all $j = 1, \cdots, p$.

2: Define $\hat{\eta}_i = \hat{\alpha} + \sum_{j=1}^{p} \hat{f}(x_{i,j})$ and $\hat{p}_i = 1/[1 + \exp(-\hat{\eta}_i)]$. Iterate

i. Construct the working target variable

$$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)};$$

ii. Construct weights $w_i = \hat{p}_i(1-\hat{p}_i)$;

iii. Fit an additive model to the targets z_i with weights w_i, using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}, \hat{f}_j$ for all $j = 1, \cdots, p$.

3: Continue the preceding step until the change in the functions falls below a pre-specified threshold.

7. Comments on Generalized Additive Models:

(a) *Advantage:* Generalized additive models provide a useful extension of linear models, making them more flexible while still retaining much of their interpretability.

(b) *Disadvantage:* The backfitting algorithm fits *all* predictors, which is *not* feasible or desirable when a large number are available.

II. Sparse Additive Models

1. Motivation and Assumption: Assume the response variable has zero mean, i.e., $\mathbb{E}[Y] = 0$. Motivated by sparsity, we assume that there exists a strictly proper, but unknown, subset $S \subset \{1, 2, \cdots, p\}$ such that the regression function can be approximated

by the sum of f_j's for $j \in S$ exclusively; that is,

$$\mathbb{E}[Y \mid X_1, X_2, \cdots, X_p] = \sum_{j \in S} f_j(X_j).$$

2. **(Naive) Population-version Problem Formulation:** For a given sparsity level $k \subset \{1, 2, \cdots, p\}$, the *best k-sparse approximation* to the regression function is given by

$$\underset{f_j \in \mathcal{F}_j \text{ for all } j=1,2,\cdots,p, |S|=k}{\arg\min} \left\{ \frac{1}{2} \mathbb{E}\left[\left(Y - \sum_{j \in S} f_j(X_j) \right)^2 \right] \right\},$$

where $\mathcal{F}_j$ is some pre-specified function class, for all $j = 1, 2, \cdots, p$.

Comment: This preceding criterion is non-convex and computationally intractable, due to combinatorial number – namely $\binom{p}{k}$ – of possible subsets of size k.

3. **Convex Population-version Problem Formulation:** We measure the sparsity of the approximation $f = \sum_{j=1}^p f_j$ by

$$\sum_{j=1}^p \|f_j\|_2,$$

where

$$\|f_j\|_2 := \sqrt{\mathbb{E}[f_j^2(X_j)]}$$

is the L_2 norm applied to the j-th component. For a given regularization parameter $\lambda \geq 0$, we minimize the following penalized criterion

$$\underset{f_j \in \mathcal{F}_j \text{ for all } j=1,2,\cdots,p}{\text{minimize}} \left\{ \frac{1}{2} \mathbb{E}\left[\left(Y - \sum_{j=1}^p f_j(X_j) \right)^2 \right] + \lambda \sum_{j=1}^p \|f_j\|_2 \right\}. \tag{7}$$

4. **Solution to (7) and Sparse Backfitting Equations:** Let $(\tilde{f}_1, \tilde{f}_2, \cdots, \tilde{f}_p)$ be the minimizer of (7). Then, for all $j = 1, 2, \cdots, p$,

$$\tilde{f}_j = \left[1 - \frac{\lambda}{\|P_j\|_2} \right]_+ P_j, \tag{8}$$

where $[x]_+ = \max\{x, 0\}$ denotes the positive part of x,

$$P_j := \mathbb{E}\left[Y - \sum_{k \neq j} \tilde{f}_k(X_k) \,\middle|\, X_j \right]$$

is the projection of the residual $R_j := Y - \sum_{k \neq j} \tilde{f}_k$ onto $\mathcal{F}_j$, and $\|P_j\|_2 = \sqrt{\mathbb{E}[P_j^2]}$. Equation (8) is called the *sparse backfitting equation*.

5. **Backfitting Algorithm for Sparse Additive Models:** Suppose we have data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, satisfying $\sum_{i=1}^n y_i = 0$. The population-version sparse backfitting equations above motivate the following sample-version sparse backfitting equations

$$\widehat{P}_j \leftarrow \mathcal{S}_j\left[\left\{y_i - \sum_{k\neq j}\hat{f}_k(x_{i,j})\right\}_{i=1}^n\right], \tag{9}$$

$$\hat{f}_j \leftarrow \left[1 - \frac{\lambda}{\|\widehat{P}_j\|_2}\right]_+ \widehat{P}_j, \tag{10}$$

for all $j = 1, 2, \cdots, p$. We cycle through all $j = 1, 2, \cdots, p$ and iterate the updates above until convergence.

The complete algorithm is given in Algorithm 3.

Algorithm 3 Backfitting Algorithm for Sparse Additive Models

1: Initialize $\hat{f}_j = 0$ for all $j = 1, 2, \cdots, p$;
2: Iterate until convergence: For each $j = 1, \cdots, p$,

Step1: Compute the i-th residual,

$$r_{i,j} = y_i - \sum_{k\neq j} f_j(x_{i,j}), \qquad \text{for all } i = 1, 2, \cdots, n;$$

Step 2: Estimate $P_j = \mathbb{E}[R_j \,|\, X_j]$ by smoothing

$$\widehat{P}_j = \mathcal{S}_j\big(\{r_{i,j}\}_{i=1}^n\big);$$

Step 3: Estimate the norm $\|P_j\|_2$ by

$$\|\widehat{P}_j\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^p (\widehat{P}_j(x_{i,j}))^2}$$

Step 4: Soft-threshold

$$\hat{f}_j = \left[1 - \frac{\lambda}{\|\widehat{P}_j\|_2}\right]_+ \widehat{P}_j; \tag{11}$$

Step 5: Center

$$\hat{f}_j = \hat{f}_j - \frac{1}{n}\sum_{i=1}^n \hat{f}_j(x_{i,j});$$

3: Output component functions $\hat{f}_j$ and the estimator $\sum_{i=1}^p \hat{f}_j$.

Remark. This algorithm can be seen as a functional version of the coordinate descent algorithm for solving the lasso problem. In particular, if we solve the lasso by iteratively minimizing with respect to a single coordinate, each iteration is given by the soft-thresholding operator.

6. **Sparse Additive Model via Basis Functions:** Let $\{\varphi_{j,k}\}_{k=1}^{\infty}$ be a set of orthonormal basis functions for $\mathcal{H}_j$ associated with the variable X_j. We can express f_j as

$$f_j = \sum_{k=1}^{\infty} \beta_{j,k}\varphi_{j,k}, \tag{12}$$

where $\beta_{j,k} = \int f_j(x_j)\varphi_{j,k}(x_j)\mathrm{d}x_j$.

Also, consider the following truncated sum of (12)

$$\tilde{f}_j := \sum_{k=1}^{d_j} \beta_{j,k}\varphi_{j,k}$$

where we only retain the first d terms and ignore the remaining ones, and the choice of d may depend on n.

Under this setup, the smoother can be taken to be the least squares projection onto the truncated set of basis functions $\{\varphi_{j,1}, \cdots, \varphi_{j,d}\}_{j=1}^{p}$. Let $\mathbf{\Phi}_j$ denote the $n \times d$ matrix whose (i,k)-entry is given by $\varphi_{j,k}(x_{i,j})$. In this case, the backfitting algorithm 3 is a coordinate descent algorithm for minimizing

$$\frac{1}{2n}\left\|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{\Phi}_j\boldsymbol{\beta}_j\right\|_2^2 + \lambda\sum_{j=1}^{p}\sqrt{\frac{1}{n}\boldsymbol{\beta}_j^{\top}\mathbf{\Phi}_j^{\top}\mathbf{\Phi}_j\boldsymbol{\beta}_j}, \tag{13}$$

where $\boldsymbol{\beta}_j := (\beta_{j,1}, \beta_{j,2}, \cdots, \beta_{j,d})^{\top} \in \mathbb{R}^d$.

Remarks.

(a) The formulation (13) is the sample version of (7).

(b) The formulation (13) is the Lagrangian of a second-order cone program, and standard convexity theory implies the existence of a minimizer.

7. **Connections with the Grouped Lasso:** The sparse additive model can be thought of as a functional version of the grouped lasso.

Review of the Grouped Lasso: The grouped lasso solves the following optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2}\left\|\mathbf{Y} - \sum_{\ell=1}^{L}\mathbf{X}_\ell\boldsymbol{\beta}_\ell\right\|_2^2 + \lambda\sum_{\ell=1}^{L}\sqrt{p_\ell}\|\boldsymbol{\beta}_\ell\|_2\right\}, \tag{14}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is the response vector, L denotes the number of groups of variables, p_ℓ denotes the number of variables in the ℓ-th group and satisfies $\sum_{\ell=1}^{L} p_\ell = p$, $\mathbf{X}_\ell \in \mathbb{R}^{n\times p_\ell}$

denotes the design matrix of the ℓ-th group of variables, $\boldsymbol{\beta}_\ell \in \mathbb{R}^{p_\ell}$ is the corresponding coefficient vector, and $\lambda \geq 0$ is the penalty parameter. We assume $\mathbf{Y}$ and each column of $\mathbf{X}$ have been centered, and $\mathbf{X}_\ell^\top \mathbf{X}_\ell = \mathbf{I}_{p_\ell}$. In addition, let $\mathbf{X} := (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L) \in \mathbb{R}^{n \times p}$ be the complete design matrix and $\boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \cdots, \boldsymbol{\beta}_L^\top)^\top \in \mathbb{R}^p$ be the complete coefficient vector.

The Karush-Kuhn-Tucker optimality conditions for the grouped lasso are

$$\begin{aligned} -\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \frac{\lambda\sqrt{p_\ell}\widehat{\boldsymbol{\beta}}_\ell}{\|\widehat{\boldsymbol{\beta}}_\ell\|_2} = \mathbf{0}_{p_\ell}, \qquad & \text{for all } \widehat{\boldsymbol{\beta}}_\ell \neq \mathbf{0}_{p_\ell}, \\ \|\mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\| \leq \lambda\sqrt{p_\ell}, \qquad & \text{for all } \widehat{\boldsymbol{\beta}}_\ell = \mathbf{0}_{p_\ell}. \end{aligned}$$

A solution to the KKT conditions above satisfies

$$\hat{\boldsymbol{\beta}}_\ell = \left[1 - \frac{\lambda\sqrt{p_\ell}}{\|\mathbf{S}_\ell\|_2}\right]_+ \mathbf{S}_\ell, \tag{15}$$

where $\mathbf{S}_\ell := \mathbf{X}_\ell^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\setminus\ell})$ and $\widehat{\boldsymbol{\beta}}_{\setminus\ell} := (\widehat{\boldsymbol{\beta}}_1^\top, \cdots, \widehat{\boldsymbol{\beta}}_{\ell-1}^\top, \mathbf{0}_{p_\ell}^\top, \widehat{\boldsymbol{\beta}}_{\ell+1}^\top, \cdots, \widehat{\boldsymbol{\beta}}_L^\top)^\top$. By iteratively applying (15), the grouped lasso solution can be obtained.

Compareing (15) with (11) reveals the similarity of the two.

III. Component Selection and Smoothing Operator

1. **Assumptions:** Let $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$ be the data. Assume $\sum_{i=1}^n y_i = 0$ and each predictor X_j has been rescaled to the unit interval, i.e., $x_{i,j} \in [0,1]$ for all $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

2. **Problem Formulation:** The *component selection and smoothing operator*, or COSSO for short, is based on the following minimization problem

$$\underset{f_j \in \mathcal{H}_j \text{ for all } j=1,2,\cdots,p}{\text{minimize}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{j=1}^p f_j(x_{i,j})\right)^2 + \tau \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j} \right\}, \tag{16}$$

where each $\mathcal{H}_j$ is a reproducing kernel Hilbert space (RKHS) of functions over $[0,1]$, and, for all $j = 1, 2, \cdots, p$,

$$\|f\|_{\mathcal{H}_j}^2 := \left(\int_0^1 f(t)\mathrm{d}t\right)^2 + \left(\int_0^1 f'(t)\mathrm{d}t\right)^2 + \int_0^1 (f''(t))^2\mathrm{d}t, \qquad \text{for all } f \in \mathcal{H}_j. \tag{17}$$

Remark 1. Note that the norm used in the penalty term in (16) is the norm itself but not the squared norm. Such a choice encourages the sparsity. In order words, by choosing $\lambda > 0$ appropriately, there exists some $j \in \{1, 2, \cdots, p\}$ such that $f_j = 0$.

Remark 2. With the choice of (17) as the norm, the associated RKHS is the space of cubic splines over $[0,1]$ with knots at the sample values of each X_j.

3. **Characterizing the Solution to (16):** Let $K_j : [0,1] \times [0,1] \to \mathbb{R}$ be the kernel function associated with $\mathcal{H}_j$, and $\mathbf{K}_j \in \mathbb{R}^{n \times n}$ be the Gram matrix whose (i,ℓ)-th entry being $K_j(x_{i,j}, x_{\ell,j})$. Then, the solution to (16) can be shown to be of the form

$$f_j = \sum_{i=1}^{n} \theta_{j,i} K_j(x_{i,j}, \cdot), \qquad \text{for all } j = 1, 2, \cdots, p.$$

It then follows that (16) can be written as

$$\underset{\boldsymbol{\theta}_j \in \mathbb{R}^n \text{ for all } j=1,2,\cdots,p}{\text{minimize}} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \sum_{j=1}^{p} \mathbf{K}_j \boldsymbol{\theta}_j \right\|_2^2 + \tau \sum_{i=1}^{p} \sqrt{\boldsymbol{\theta}_j^\top \mathbf{K}_j \boldsymbol{\theta}_j} \right\}, \tag{18}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the response vector and $\boldsymbol{\theta}_j := (\theta_{j,1}, \theta_{j,2}, \cdots, \theta_{j,n})^\top \in \mathbb{R}^n$ for all $j = 1, 2, \cdots, n$.

Remark 1. Even though the original problem (16) is of infinite dimensional, the solution for each coordinate resides in a finite dimensional subspace. Hence, collectively, the solution to (16) is of finite dimensional.

Remark 2. Note that (18) is very similar to the objective function in the grouped lasso problem (14).

4. **Optimality Conditions:** Let $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \cdots, \widehat{\boldsymbol{\theta}}_p)$ be the minimizer of (18). Then, it satisfies

$$\widehat{\boldsymbol{\theta}}_j = \begin{cases} \mathbf{0}_n, & \text{if } \sqrt{\mathbf{r}_j^\top \mathbf{K}_j \mathbf{r}_j} < \tau, \\ \left(\mathbf{K}_j + \frac{\tau}{\sqrt{\widehat{\boldsymbol{\theta}}_j^\top \mathbf{K}_j \widehat{\boldsymbol{\theta}}_j}} \mathbf{I}_n \right)^{-1} \mathbf{r}_j, & \text{otherwise,} \end{cases}$$

where $\mathbf{r}_j := \mathbf{Y} - \sum_{\ell \neq j} \mathbf{K}_\ell \widehat{\boldsymbol{\theta}}_\ell$ corresponds to the j-th partial residual.

Remark. An algorithm to compute $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \cdots, \widehat{\boldsymbol{\theta}}_p)$ can be obtained by using the spectral decomposition of $\mathbf{K}_j$, for all $j = 1, 2, \cdots, n$, and a simple one-dimensional search.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN: 1498712169.