Segmentation of Counting Processes and Dynamical Models

Mokhtar Z. Alaya

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

THÈSE DE DOCTORAT

en vue de l'obtention du grade de

Docteur ès Sciences de l'Université Pierre et Marie Curie

Discipline : Mathématiques

Spécialité : Statistique

présentée par

Mokhtar Zahdi ALAYA

Segmentation of Counting Processes and Dynamical Models

---

dirigée par Stéphane GAÏFFAS et par Agathe GUILLOUX

Au vu des rapports établis
par MM. Jacobo DE UÑA-ÁLVAREZ et Erwan LE PENNEC

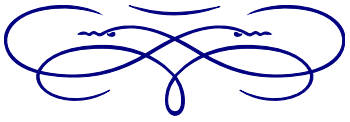
Soutenue le 27 juin 2016 devant le jury composé de :

M. Pierre ALQUIER	ENSAE	Examinateur
M. Sylvain ARLOT	Université Paris-Sud	Examinateur
M. Gérard BIAU	Université Pierre et Marie Curie	Examinateur
M. Stéphane GAÏFFAS	École Polytechnique	Directeur de thèse
M^{me} Agathe GUILLOUX	Université Pierre et Marie Curie	Directrice de thèse
M. Erwan LE PENNEC	École Polytechnique	Rapporteur

Laboratoire de Statistique Théorique
et Appliquée (LSTA)
4, place Jussieu
75252 Paris Cedex 05

Sorbonnes Universités, UPMC
École Doctorale de Sciences
Mathématiques de Paris Centre
4, place Jussieu
75252 Paris Cedex 05

Dédicace



À ma mère Sahara ; celle qui a passé sa vie les mains levées, me souhaitant la grâce et la bénédiction.

Remerciements

Mes premiers remerciements s'adressent à mes directeurs Stéphane et Agathe.

Stéphane et Agathe, je vous remercie pour la chaleur de votre accueil et pour la confiance que vous m'avez témoignée pendant ces années de thèse. Je vous suis reconnaissant de m'avoir proposé un sujet aussi riche et passionnant. Merci pour votre disponibilité, vos nombreux conseils, votre soutien permanent, et vos encouragements dans les périodes de doute.

Je suis très honoré que Jacobo De Uña-Álvarez et Erwan Le Pennec aient accepté d'être rapporteurs de ma thèse. Je leur suis très reconnaissant du temps qu'ils ont consacré à l'évaluation de ce travail. Je remercie également Pierre Alquier, Sylvain Arlot et Gérard Biau d'avoir accepté de faire partie de mon jury.

Je tiens à exprimer toute mon amitié aux membres du LSTA que j'ai côtoyés pendant ces quatre années. Merci pour votre sympathie, votre bonne humeur et pour l'ambiance agréable de tous les jours. Je remercie tout particulièrement Michel Broniatowski pour les conseils qu'il m'a donnés.

J'adresse mes remerciements à tous mes collègues docteur-e-s et doctorant-e-s : Ahmed, Amadou, Assia, Baptiste, Benjamin, Boris, Cécile, Diaa, Erwan, Émilie, Félix, Layal, Lucie, Malamine, Matthieu, Mohamed, Monia, Nazih, Nedjemeddine, Quyen, Roxane, Sarah, Simon, Soumeya, Svetlana, Tarn, Thibault, Zaid, Zeineb pour tous les bons moments passés ensemble. Bon courage à ceux qui terminent bientôt.

Toute ma reconnaissance va aussi à ma famille qui m'a énormément soutenu, notamment à mon frère Radhouane et sa femme Kmar, et à mes frères et mes sœurs en Tunisie, loin de moi mais très présents par leurs pensées et leurs vifs encouragements. Sachez que je ne saurai vous rendre ce que vous m'avez apporté, merci pour tout et que Dieu vous garde.

Je terminerai par ma femme Sarra. Je ne pourrai jamais assez te remercier pour ton soutien sans faille et indéfectible. Les mots ne suffisent pas à exprimer à quel point je te suis reconnaissant pour ta patience et tes encouragements quotidiens.

Avant-propos

Cette thèse a été financée par un contrat doctoral à l'Université Pierre et Marie Curie, du 1er octobre 2012 au 30 septembre 2015, et réalisée au Laboratoire de Statistique Théorique et Appliquée (LSTA). Elle est composée de 4 chapitres : les chapitres 2 à 4 peuvent être lus indépendamment les uns des autres.

Plan de la thèse

Le Chapitre 1 est dédié d'une part à présenter les concepts fondamentaux utilisés dans ce travail, et d'autre part à synthétiser les principaux résultats. Nous introduisons tout d'abord l'apprentissage statistique pour le modèle simple de régression linéaire. Nous présentons ensuite quelques méthodes d'estimation de l'intensité d'un processus de comptage. Nous finissons par une vue d'ensemble de nos contributions :

— l'apprentissage pour l'intensité d'événements récurrents avec points de rupture,
— la prédiction en grande dimension avec une sparsité induite par la binarisation de variables, et
— l'estimation de paramètres dans les modèles d'Aalen et de Cox avec covariables en grande dimension et dépendant du temps.

Le Chapitre 2 fait l'objet d'un article publié dans IEEE, Transactions on Information Theory, écrit en collaboration avec Stéphane Gaïffas (CMAP - École Polytechnique) et Agathe Guilloux (LSTA - UPMC). Dans ce chapitre, nous introduisons une procédure d'estimation basée sur la pénalisation par variation totale avec poids, permettant une calibration fine de la relaxation convexe de l'hypothèse considérée. Nous proposons des inégalités oracles exactes pour cette procédure avec une vitesse rapide de convergence, et nous démontrons la consistance de cette méthode pour la détection des points de rupture. Ces résultats fournissent ainsi une première garantie théorique pour la segmentation basée sur une relaxation convexe au delà du cadre signal + bruit blanc gaussien habituellement étudiée. Nous introduisons un algorithme efficace pour résoudre le problème convexe sous-jacent et nous illustrons notre approche sur des données simulées et des données génomiques.

Le Chapitre 3 fait l'objet d'un travail écrit en collaboration avec Stéphane Gaïffas et Agathe Guilloux. Nous nous intéressons à la construction et à la mise en œuvre d'une nouvelle notion de régularisation nommée "binarsity". Elle compte le nombre de valeurs différentes du vecteur de paramètres à estimer dans un espace engendré par des variables binarisées. Nous considérons une procédure d'estimation basée sur une version data-driven

de binarsity de la fonction de régression d'un modèle linéaire généralisé. Nous fournissons une garantie théorique non asymptotique des performances de l'estimateur donné, et un algorithme proximal pour la résolution du problème convexe étudié.

Le Chapitre 4 fait l'objet d'un travail écrit en collaboration avec Thibault Allart (CNAM - Ubisoft), Agathe Guilloux, et Sarah Lemler (École Centrale Supélec). Nous considérons le problème d'estimation des coefficients de régression dans les modèles Alaen et Cox quand les coefficients et les covariables (en grande dimension) dépendent du temps. Pour cela, nous utilisons une procédure d'estimation spécifique basée sur la pénalisation par variation totale. Nous donnons des inégalités oracles pour les estimateurs proposés, et nous nous intéressons ensuite à la résolution algorithmique de ces estimateurs par des méthodes proximales en optimisation convexe.

Table des matières

Chapitre 1

État de l'Art et Contributions

Sommaire

L'enjeu statistique des travaux présentés dans ce manuscrit s'articule autour de trois problématiques : la segmentation de l'intensité d'un processus de comptage, la binarisation de variables explicatives continues dans les modèles linéaires généralisés, et les modèles de régression dynamique avec coefficients et covariables dépendant du temps. Pour ces trois problèmes, nous adoptons des techniques d'apprentissage statistique supervisé en grande dimension. Nous introduisons des procédures d'estimation basées sur la pénalisation par variation totale avec poids. Nous étudions les propriétés théoriques des estimateurs, et nous décrivons des algorithmes permettant de les calculer. Dans ce chapitre, nous présentons un état de l'art non exhaustif concernant l'apprentissage statistique supervisé linéaire et quelques outils d'optimisation convexe.

1.1 Apprentissage statistique supervisé

On suppose dans la suite que l'on a accès à un échantillon $\mathcal{D}_n$ (également appelée échantillon d'apprentissage) composé de n couples d' entrées-sorties (X_i, Y_i) tel que $X_i \in \mathcal{X} \subset \mathbb{R}^p$

(variable explicative) et $Y_i \in \mathscr{Y} \subset \mathbb{R}$ (variable réponse) pour $i = 1,\dots,n$. Ces couples de variables sont indépendantes et identiquement distribuées, et indépendantes et de même loi que le couple (X,Y). Nous notons par $\boldsymbol{X} = [X_1,\dots,X_n]^\top$ la matrice des variables explicatives (prédicteurs) et $\boldsymbol{Y} = [Y_1,\dots,Y_n]^\top$ le vecteur de $\mathscr{Y}^n$ contenant les labels de l'ensemble d'apprentissage. Le contexte de grande dimension se traduit par le fait que le nombre p des variables explicatives est grand par rapport au nombre n des réponses.

L'objectif de l'apprentissage supervisé est la construction d'un prédicteur $h : \mathscr{X} \to \mathbb{R}$ (ou classifieur si on s'intéresse au problème de classification), capable de prédire un label y' pour une nouvelle observation x'. Par souci de simplicité, nous supposons dans cette section que h est une fonction linéaire de la forme $h(x) = h_\beta(x) = x^\top\beta$, avec $\beta \in \mathbb{R}^p$.

1.1.1 Minimisation du risque empirique pénalisé

Pour mesurer la qualité de prédiction, nous introduisons une fonction de coût ou fonction de perte $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$. Elle mesure l'adéquation du prédicteur $x^\top\beta$ aux données observées. Ainsi, $\ell(y, x^\top\beta)$ quantifie l'écart entre la sortie y associée à une variable x et sa prédiction $x^\top\beta$. La détermination du prédicteur $X^\top\beta$ revient à résoudre un problème d'optimisation. En effet, le meilleur prédicteur possible est celui qui minimise l'espérance de l'erreur de prédiction, aussi appelé erreur de généralisation

$$R(\beta) = \mathbb{E}[\ell(Y, X^\top\beta)] = \int_{\mathscr{X}\times\mathscr{Y}} \ell(y, x^\top\beta) d\mathbb{P}_{(X,Y)}(x,y).$$

Ce risque mesure la capacité de généralisation du prédicteur $X^\top\beta$. Il ne peut cependant pas être minimisé en pratique car la loi jointe $\mathbb{P}_{(X,Y)}$ est inconnue, ce qui rend la minimisation de $R(\beta)$ impossible. L'idée est de remplacer la mesure théorique par sa mesure empirique $d\mathbb{P}_n(x,y) = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}(x,y)$. Nous cherchons donc à minimiser le risque empirique (Vapnik (1995, 1998)) défini par

$$\underset{\beta\in\mathbb{R}^p}{\text{minimise}} \Big\{ R_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) \Big\}.$$

La minimisation du risque empirique peut conduire à des instabilités numériques et à de mauvaises performances en généralisation. En effet, les algorithmes employant ce principe sont souvents sujets au sur-apprentissage. Le sur-apprentissage a lieu lorsque le prédicteur $X^\top\beta$ appris ne commet quasiment aucune erreur sur les données d'apprentissage (risque empirique très faible) mais commet beaucoup d'erreurs sur les autres données (erreur de généralisation très élevé).

Pour éviter le sur-apprentissage, nous pouvons ajouter au risque empirique un terme de régularisation $\lambda J(\beta)$, avec $\lambda \geq 0$ et $J(\beta) : \mathbb{R}^p \to \mathbb{R}_+$ une fonction prenant en compte la complexité du modèle. Ceci conduit ainsi à la minimisation du risque empirique pénalisé (Evgeniou et al. (2002); Vapnik (1998)) défini par

$$\underset{\beta\in\mathbb{R}^p}{\text{minimise}} \big\{ R_n(\beta) + \lambda J(\beta) \big\}. \tag{1.1}$$

Le coefficient λ, appelé paramètre de régularisation, contrôle le compromis entre l'adéquation du modèle aux données et la complexité du modèle.

Il existe de nombreux types de régularisations parmi lesquels le coefficient C_p de Mallows (1973), le critère AIC d' Akaike (1974), le critère BIC de Schwarz (1978), le Lasso de Tibshirani (1996), la sélection de modèles de Birgé and Massart (2001, 2007), parmi beaucoup d'autres. La famille de régularisations majoritairement utilisée en apprentissage supervisé est celle des normes et pseudo-normes ℓ_q

$$J_q(\beta) = \|\beta\|_q = \Big(\sum_{j=1}^{q} |\beta_j|^q\Big)^{1/q},$$

avec $0 \leq q \leq +\infty$. Notons que les pseudo-normes ℓ_q pour $q \in [0,1)$ ne sont pas convexes comme ci-illustré dans la Figure 1.1.

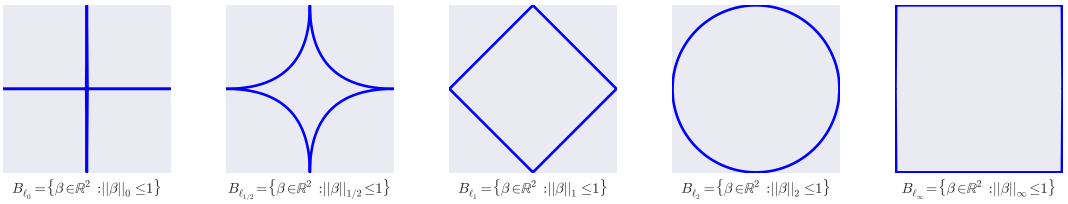


Fig. 1.1 – Boule unité de $\mathbb{R}^2$ pour les normes $\ell_0, \ell_{1/2}, \ell_1, \ell_2$, et $\ell_{+\infty}$.

1.1.2 Lasso et ses dérivées

Plusieurs approches basées sur la pénalisation par les (pseudo) normes ℓ_q ont été introduites pour le modèle de régression linéaire simple. Nous présentons les plus populaires d'entre elles pour ce modèle qui s'écrit $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ où ε est un bruit à variance supposée ici connue.

Pseudo norme ℓ_0. La pseudo-norme ℓ_0 consiste à compter le nombre d'éléments non nuls d'un vecteur, c'est à dire pour tout $\beta \in \mathbb{R}^p$

$$\|\beta\|_0 = \big|\{j : \beta_j \neq 0, j = 1, \ldots, p\}\big| = \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0) = \lim_{q \to 0} \|\beta\|_q^q.$$

Le problème de minimisation du risque empirique pénalisé par la pseudo-norme ℓ_0 est défini par

$$\hat{\beta}_{\text{BIC}} = \text{argmin}_{\beta \in \mathbb{R}^p} \Big\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_0 \Big\}. \tag{1.2}$$

La pénalité ℓ_0 a été parmi les premières étudiées (Schwarz (1978)). L'inconvénient majeur de ce cette dernière est du au fait que le problème de minimisation (1.2) est non convexe et NP-Dur, c'est à dire qu'il n'est pas résolvable en un temps polynomial, voir Natarajan (1995) et Tropp (2004). Pour palier cet inconvénient d'un point vue algorithmique, plusieurs auteurs ont proposé de convexifier le problème d'optimisation (1.2). C'est l'approche qui prévaut dans la construction des estimateurs Lasso et ses dérivées.

Lasso (norme ℓ_1). La norme ℓ_1 fut introduite par Tibshirani (1996) et est plus connue sous le nom du Lasso (Least absolute shrinkage and selection operator). Cette pénalité est

utilisée pour effectuer une sélection automatique des variables. Le problème Lasso s'écrit sous la forme

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\beta\|_2^2+\lambda\|\beta\|_1\right\}. \tag{1.3}$$

Le choix de λ est crucial. En effet, pour une très grande valeur de λ et donc une très forte pénalité il ne reste aucune variable considérée comme pertinente. Lorsque λ diminue, le nombre de variables pertinentes augmente jusqu'à atteindre le modèle maximal pour $\lambda = 0$. Il existe une importante littérature sur l'estimateur Lasso : Knight and Fu (2000) établissent la consistance de l'estimateur Lasso pour un paramètre de régularisation spécifique $\lambda = \lambda_n$. Sous une hypothèse d'irreprésentabilité, Zhao and Yu (2006) démontrent que le Lasso est signe-consistant, c'est à dire l'estimateur $\hat{\beta}_{\text{Lasso}}$ possède asymptotiquement les mêmes signes que le vrai paramètre β. D'autres propriétés théoriques de modèle (1.3) sont étudiées par Bickel et al. (2009); Bühlmann and Van De Geer (2011); Bunea et al. (2007); Candes and Tao (2007); Meinshausen and Yu (2009); Tibshirani (2013); van de Geer and Bühlmann (2009); Wainwright (2009); Zhang (2009). Nous renvoyons à Hebiri pour une présentation détaillée des résultats existants sur le Lasso. Algorithmiquement, le Lasso est solution d'un problème de minimisation convexe. De nombreux algorithmes efficaces convergent rapidement vers cette solution. L'un des plus populaires est le Lars (least angle regression stepwise), proposé par Efron et al. (2004), et qui permet d'obtenir le chemin de régularisation de l'estimateur Lasso pour une plage de valeurs de λ.

Adaptive Lasso (norme ℓ_1 avec poids). L'adaptive Lasso étudié par Zou (2006) est une version modifiée du Lasso. Il procède en deux temps : le statisticien calcule tout d'abord un estimateur initial $\hat{\beta}_{\text{init}} \in \mathbb{R}^p$, pouvant être l'estimateur des moindres carrées, ridge, Lasso, ou tout autre estimateur. Ce premier estimateur va être utilisé comme poids dans une deuxième étape d'estimation Lasso, en fixant $\hat{w}_j = \hat{\beta}_{j,\text{init}}$ tel que :

$$\hat{\beta}_{\text{aLasso}} = \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\beta\|_2^2+\lambda\sum_{j=1}^{p}\frac{|\beta_j|}{\hat{w}_j}\right\}.$$

Chacun des poids $\hat{w}_j$ permet de doser le niveau de pénalité de chaque coefficient β_j. En effet, si $\hat{w}_j = 0$ alors $\hat{\beta}_{j,\text{aLasso}} = 0$ de telle sorte que la première étape de Lasso sert de présélection. De plus, si $\hat{w}_j$ est grand, l'adaptive Lasso utilise une pénalisation plus petite, donc un rétrécissement plus petit pour le j-ème coefficient. Zou (2006) a montré que si l'estimateur initial est consistant, l'adaptive Lasso sera lui aussi consistant. L'adaptive Lasso reste un problème convexe et les algorithmes développés pour la résolution du Lasso peuvent aisément s'adapter.

Elastic Net (norme $\ell_1 + \ell_2$). L'Elastic Net (Zou and Hastie (2005)) généralise le Lasso et construit des modèles combinant deux propriétés désirables : la sparsité assurée par la pénalité de type Lasso (norme ℓ_1) et la capture des corrélations entre les variables par la pénalité ridge (norme ℓ_2). La méthode de l'Elastic Net est définie par

$$\hat{\beta}_{\text{eNet}} = \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\beta\|_2^2+\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2\right\},$$

où λ_1 et λ_2 sont deux paramètres de régularisation. La régression Elastic Net était initialement motivée par des applications en biologie computationnelle, où les variables explicatives présentent une forte corrélation. L'estimateur Lasso n'est pas idéal pour ces modèles

compte tenu de ses difficultés pratiques créées par la présence de fortes corrélations entre les variables. L'estimateur $\hat{\beta}_{\text{eNet}}$ peut être calculé par l'algorithme Lars-eNet (Zou and Hastie (2005)). Nous citons quelques extensions de la méthode Elastic Net : Bunea (2008) l'a étendue pour la régression logistique et Li and Lin (2010) ont proposé une formulation bayésienne de l'Elastic Net. Pour obtenir des propriétés oracles de cette méthode, Zou and Zhang (2009) ont proposé l'adaptive Elastic Net.

Fused Lasso (norme $\ell_1 + \text{TV}$). La pénalisation par variation totale $\|\cdot\|_{\text{TV}}$ est définie par la norme ℓ_1 du gradient discret, i.e.,

$$\|\beta\|_{\text{TV}} = \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|,$$

pour tout $\beta \in \mathbb{R}^p$. La variation totale mesure les changements des valeurs au sein d'un signal. C'est une pénalité très efficace pour reconstruire les signaux constants par morceaux (Little and Jones (2011); Rudin et al. (1992)).

En couplant les pénalités ℓ_1 et TV, Tibshirani et al. (2005) ont construit l'estimateur fused Lasso

$$\hat{\beta}_{\text{fLasso}} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_{\text{TV}} \right\},$$

où λ_1 et λ_2 sont deux paramètres de régularisation. La pénalité ℓ_1 encourage la sparsité des coefficients, tandis que la pénalité TV encourage la sparsité de leurs différences. Cette dernière permet à l'estimateur fused Lasso de répondre à des problèmes où les variables sont ordonnées (Tibshirani et al. (2005)). D'un point de vue algorithmique, Friedman et al. (2007), Jun et al. (2010), et Hoefling (2010) proposent des algorithmes efficaces pour calculer l'estimateur $\hat{\beta}_{\text{fLasso}}$. Des résultats asymptotiques ont été étudié par Rinaldo (2009) dans le contexte de l'estimation d'un signal sparse par bloc. Le fused Lasso a été aussi utilisé dans le problème de la détection des points de rupture par Harchaoui and Lévy-Leduc (2010) et Bleakley and Vert.

Group-Lasso (norme ℓ_1/ℓ_2). La méthode de Group-Lasso (Yuan and Lin (2006)) généralise le Lasso aux variables par groupes. Décrivons alors une partition de $\{1,\dots,p\}$ par les ensembles $(G_k)_{k=1,\dots,K}$, où G_k est l'ensemble des indices des variables contenues dans le groupe G_k avec $k = 1,\dots,K$. L'estimateur Group Lasso est alors défini par

$$\hat{\beta}_{\text{gLasso}} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{k=1}^{K} \|\beta_{G_k}\|_2 \right\},$$

où $\|\beta_{G_k}\|_2 = \sqrt{\sum_{j \in G_k} \beta_j^2}$. La pénalisation group-Lasso consiste à appliquer une pénalisation ℓ_1 aux normes ℓ_2 de chaque groupe. Cela aura pour effet de favoriser la sparsité des groupes. Plusieurs travaux ont analysé les propriétés statistiques de l'estimateur $\hat{\beta}_{\text{gLasso}}$. Bach (2007) établit la consistance en sélection du group-Lasso. Sous une hypothèse sur la matrice de Gram $\boldsymbol{X}^\top \boldsymbol{X}/n$, Nardi and Rinaldo (2008) démontrent des inégalités oracles. D'autres extensions du group-Lasso ont été proposé, Meier et al. (2008) exploite la pénalité groupée avec la perte logistique, Simon et al. (2013) étudient le sparse group-Lasso, Alaiz et al. (2013) proposent le group fused Lasso.

1.1.3 Inégalités oracles

Les inégalités oracles servent à énoncer des propriétés statistiques d'estimateurs. En effet, elles lient les performances d'un estimateur réel à celles d'un autre qui s'appuie sur l'information donnée par un oracle (selon la terminologie introduite par Donoho and Johnstone (1998)). Ce dernier est le meilleur estimateur $\beta^\star \in \mathbb{R}^p$ qui minimise le risque $R(\beta), i.e.$

$$\beta^\star = \text{argmin}_{\beta\in\mathbb{R}^p} R(\beta).$$

Cependant, il dépend de la loi des données qui est inconnue. Il est donc inaccessible mais son risque peut servir de référence. Autrement dit, il est possible de construire des estimateurs qui miment les performances de l'oracle en terme de risque. Cette propriété est au cœur de l'inégalité oracle donnée par

$$R(\hat{\beta}) \leq C \inf_{\beta\in\mathbb{R}^p} \{R(\beta) + \epsilon_{n,p}(\beta)\},$$

avec une grande probabilité, ou

$$\mathbb{E}[R(\hat{\beta})] \leq C \inf_{\beta\in\mathbb{R}^p} \{R(\beta) + \epsilon_{n,p}(\beta)\}.$$

avec C est une constante que l'on souhaite proche de 1 (plus elle est proche de 1, plus on s'approche des performances de l'oracle), et $\epsilon_{n,p}(\cdot)$ est un reste que l'on souhaite minimal. La majorité des estimateurs définis dans la Section 1.1.2 vérifient des inégalités oracles : les travaux de (Bickel et al. (2009); Bunea et al. (2007); Lounici (2008); Zhang and Huang (2008)) pour le Lasso, Zou (2006) pour l'adaptive Lasso, Lounici et al. (2011) pour le group-Lasso, Li et al. (2010) pour l'adaptive Elastic Net. Généralement, il existe deux types d'inégalités oracles selon la vitesse de convergence de $\hat{\beta}$: nous distinguons inégalités oracles à vitesse lente (respectivement rapide) si la convergence est de l'ordre de $\sqrt{\log p/n}$ (respectivement $\log p/n$). La vitesse rapide est obtenue sous la garantie d'hypothèses supplémentaires : par exemple celle des valeurs propres restreintes de la matrice de Gram $\boldsymbol{X}^\top\boldsymbol{X}/n$, voir Bickel et al. (2009); Bunea et al. (2007); Meinshausen and Yu (2009); van de Geer and Bühlmann (2009).

1.1.4 Optimisation convexe : algorithme du gradient proximal

Dans cette sous section, nous présentons la notion d'algorithme proximal qui permet de résoudre les problèmes convexes de type (1.1). Les algorithmes proximaux ont suscité ces dernières années un intérêt croissant, grâce à leur facilité de mise en œuvre et leur simplicité. Ils font partie des techniques de premier ordre en optimisation convexe, et permettent de résoudre des problèmes non lisses (Bach et al. (2012); Beck and Teboulle (2009a,b); Chen et al. (2012); Combettes and Pesquet (2011); Combettes and Wajs (2005); Daubechies et al. (2004); Nesterov (2007)).

En apprentissage supervisé, les problèmes convexes pouvant être résolus par un algorithme proximal s'écrivent sous la forme

$$\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \Big\{ f(\beta) + \lambda J(\beta) \Big\}, \tag{1.4}$$

avec $f(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i^\top\beta)$ est une fonction d'attaches au données supposée convexe, différentiable et admettant un gradient L-Lipschitz. La pénalité J est aussi convexe et simple.

Nous disons qu'une fonction est simple si son opérateur proximal, voir Définition 1.1.1 plus bas, est facile à calculer numériquement.

Un algorithme itératif pour résoudre (1.4) consiste à majorer la fonction f au point courant de chaque itération. Pour ce faire, on utilise le principe de majoration minimisation avec un majorant quadratique donnée par le Lemme de Descente (Nesterov),

$$f(\beta) \leq f(\beta^{(k)}) + \langle \beta - \beta^{(k)}, \nabla f(\beta^{(k)}) \rangle + \frac{L}{2} \|\beta - \beta^{(k)}\|_2^2.$$

Alors,

$$f(\beta) + \lambda J(\beta) \leq f(\beta^{(k)}) + \langle \beta - \beta^{(k)}, \nabla f(\beta^{(k)}) \rangle + \frac{L}{2} \|\beta - \beta^{(k)}\|_2^2 + \lambda J(\beta). \tag{1.5}$$

Une idée naturelle à la vue de (1.5) consiste à appliquer un schéma du type

$$\beta^{(k+1)} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ f(\beta^{(k)}) + \langle \beta - \beta^{(k)}, \nabla f(\beta^{(k)}) \rangle + \frac{L}{2} \|\beta - \beta^{(k)}\|_2^2 + \lambda J(\beta) \right\}, \tag{1.6}$$

ainsi on s'assure que $f(\beta^{(k+1)}) + \lambda J(\beta^{(k+1)}) \leq f(\beta^{(k)}) + \lambda J(\beta^{(k)})$. La suite $(f(\beta^{(k)}) + \lambda J(\beta^{(k)}))_k$ est donc décroissante monotone. L'itération (1.6) peut être réécrite sous la forme

$$\begin{aligned} \beta^{(k+1)} &= \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{L}{2} \left(\|\beta - \left(\beta^{(k)} - \frac{1}{L} \nabla f(\beta^{(k)})\right)\|_2^2 - \|\frac{1}{L} \nabla f(\beta^{(k)})\|_2^2 \right) + \lambda J(\beta) \right\} \\ &= \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\beta - \left(\beta^{(k)} - \frac{1}{L} \nabla f(\beta^{(k)})\right)\|_2^2 + \frac{\lambda}{L} J(\beta) \right\}. \end{aligned} \tag{1.7}$$

Si $\lambda = 0$, alors dans ce cas nous pouvons chercher l'estimateur $\hat{\beta}$ par un algorithme de descente de gradient

$$\beta^{(k+1)} \leftarrow \beta^{(k)} - \frac{1}{L} \nabla f(\beta^{(k)}).$$

De plus, si la pénalité J est la fonction indicatrice δ_C d'un ensemble convexe C définie par

$$\delta_C(\beta) = \begin{cases} 0, \text{ si } \beta \in C, \\ +\infty, \text{ sinon,} \end{cases}$$

alors nous résolvons le problème (1.7) par un algorithme de gradient projeté avec une projection sur le convexe C. Cela motive la définition de l'opérateur proximal associé au terme de régularisation λJ.

Definition 1.1.1. (Opérateur proximal (Moreau (1962))). Soit $J : \mathbb{R}^p \to \mathbb{R}$ une fonction convexe, fermée et propre (J ne prend pas la valeur $-\infty$ et elle n'est pas identiquement égale à $+\infty$). L'opérateur proximal associé à (λ, J) noté par $\text{prox}_{\lambda J} : \mathbb{R}^p \to \mathbb{R}^p$ est défini par

$$\text{prox}_{\lambda J}(v) = \text{argmin}_{u \in \mathbb{R}^p} \left\{ \frac{1}{2} \|v - u\|_2^2 + \lambda J(u) \right\}.$$

Notons que la fonction à minimiser dans la définition de l'opérateur proximal est strictement convexe et possède donc un unique minimum pour tout $v \in \mathbb{R}^p$. L'opérateur proximal est une extension de la notion de projection, car en prenant comme cas particulier la fonction indicatrice d'un convexe fermé non vide C, on retrouve ($\text{prox}_{\delta_C}(v) = \text{argmin}_{u \in C} \|v - u\|_2$) la projection euclidienne sur C. Lorsque la pénalité J est séparable, c'est à dire $J(\beta) = \sum_{j=1}^p J_j(\beta_j)$, alors le calcul de l'opérateur proximal se fait coordonnée par coordonnée, $\left(\text{prox}_J(\beta)\right)_j =$

$\text{prox}_{J_j}(\beta_j)$. Une condition suffisante pour que l'opérateur proximal d'une somme de deux pénalités J_1 et J_2 soit égal à la composition de ses deux opérateurs proximaux est donnée par le Théorème 1 dans Yu (2013). Ce théorème nous dit que

$$\text{si } \partial(J_2(\beta)) \subset \partial(J_1(\text{prox}_{J_2}(\beta))) \text{ alors } \text{prox}_{J_1+J_2} = \text{prox}_{J_1} \circ \text{prox}_{J_2},$$

où $\partial(J(\beta))$ est la sous différentielle de la pénalité J au point β (Boyd and Vandenberghe (2004)). De nombreuses formes explicites d'opérateurs proximaux sont répertoriées dans de récents travaux, nous citons (Chaux et al. (2007); Combettes and Pesquet (2011); Combettes and Wajs (2005); Combettes and Pesquet (2007); Friedman et al. (2007)). Dans le Tableau 1.1, nous en rappelons quelques unes. Nous soulignons le fait qu'il n'existe pas de forme explicite dans le cas de la pénalisation TV. En revanche, nous utilisons dans nos travaux l'algorithme proposé par Condat (2013) pour calculer $\text{prox}_{\lambda\|\cdot\|_{\text{TV}}}$.

Pénalité	Opérateur proximal
$J(\beta) = \|\beta\|_1$	$\left(\text{prox}_{\lambda J}(\beta)\right)_j = \max\left(0, 1 - \frac{\lambda}{\|\beta_j\|}\right)\beta_j.$
$J(\beta) = \|\beta\|_2$	$\text{prox}_{\lambda J}(\beta) = \max\left(0, 1 - \frac{\lambda}{\|\beta\|_2}\right)\beta.$
$J(\beta) = \frac{1}{2}\|\beta\|_2^2$	$\text{prox}_{\lambda J(\beta)}(\beta) = \frac{1}{1+\lambda}\beta.$
$J(\beta) = \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|_2^2$	$\text{prox}_J(\beta) = \frac{1}{1+\lambda_2}\text{prox}_{\lambda_1\|\cdot\|_1}(\beta).$
$J(\beta) = \sum_{k=1}^{K}\|\beta_{G_k}\|_2$	$\left(\text{prox}_{\lambda J}(\beta)\right)_{G_k} = \max\left(0, \left(1 - \frac{\lambda}{\|\beta\|_2}\right)\right)\beta_{G_k}$
$J(\beta) = \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_{\text{TV}}$	$\text{prox}_J(\beta) = \text{prox}_{\lambda_1\|\cdot\|_1}\left(\text{prox}_{\lambda_2\|\cdot\|_{\text{TV}}}(\beta)\right)$

TABLE 1.1 – Quelques formes explicites d'opérateurs proximaux.

L'algorithme de gradient proximal repose sur l'itération suivante

$$\beta^{(k+1)} \leftarrow \text{prox}_{\frac{\lambda}{L}J}\left(\beta^{(k)} - \frac{1}{L}\nabla f(\beta^{(k)})\right).$$

Nous présentons ici la procédure ISTA (Iterative Shrinkage Thresholding Algorithm), voir Daubechies et al. (2004) et Combettes and Wajs (2005). Sous l'hypothèse que la fonction d'attache aux données f soit différentiable avec un gradient L-Lipschitz, et convexe (respectivement α-fortement convexe (Combettes and Pesquet (2011)), le taux de convergence de ISTA est $O(1/k)$ (respectivement $O\left((1-\frac{\alpha}{L})^k\right)$ (cf. Daubechies et al. (2004)). Le pseudo code de ISTA est défini dans l'Algorithme 1.

Algorithm 1: Procédure ISTA

1. Calculer la constante de Lipschitz L de l'opérateur ∇f.
2. Initialisation : $\beta^{(0)} \in \mathbb{R}^p$;
3. **repeat**
   $\beta^{(k+1)} \leftarrow \text{prox}_{\frac{\lambda}{L}J}\left(\beta^{(k)} - \frac{1}{L}\nabla\ell(\beta^{(k)})\right)$;
   until *convergence*
4. **return** β

En se basant sur des résultats élaborés par Nesterov (2007), la procédure FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) développée dans les travaux de Beck and Teboulle (2009a) accélère la vitesse de convergence de la procédure ISTA. En effet, la vitesse de

convergence de FISTA est $O(1/k^2)$ (et $O\big((1-\sqrt{\frac{\alpha}{L}})^k\big)$ si la fontion f est α-fortement convexe) (voir Beck and Teboulle (2009a). Le pseudo code de FISTA est donné dans l'Algorithme 2.

Algorithm 2: Procédure FISTA

1. Calculer la constante de Lipschitz L de l'opérateur ∇f.
2. Initialisation : $\beta^{(0)} \in \mathbb{R}^p$; $\mu^{(0)} = \beta^{(0)}$; et $t_1 = 1$;
3. **repeat**
   $\beta^{(k)} \leftarrow \text{prox}_{\frac{\lambda}{L}J}\big(\mu^{(k)} - \frac{1}{L}\nabla\ell(\mu^{(k)})\big)$;
   $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
   $\mu^{(k+1)} \leftarrow \beta^{(k)} + \big(\frac{t_k-1}{t_{k+1}}\big)(\beta^{(k)} - \beta^{(k-1)})$;
   until *convergence*
4. **return** β

1.2 Inférence statistique pour un processus de comptage

Les processus de comptage occupent une place privilégiée pour la modélisation de nombreux événements récurrents. On rencontre ces événements dans divers domaines, par exemple en génomique, en biologie, ou en économétrie (Andersen et al. (1993); Brémaud (1981); Fleming and Harrington (1991); Martinussen and Scheike (2007)). Pour décrire la dynamique d'un processus de comptage, on utilise sa fonction d'intensité. Elle est définie par la probabilité instantanée qu'un événement survienne à un instant donné conditionnellement à l'histoire du processus. Dans cette section, nous présentons le cadre mathématique des processus de comptage.

1.2.1 Processus de comptage : définitions et description du modèle

Fixons quelques notations qui serviront pour la suite. On dispose d'un espace de probabilité filtré $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,\tau]}, \mathbb{P})$ où $[0,\tau]$ est un intervalle de temps continu avec $0 \leq \tau < +\infty$ le temps terminal. La filtration $(\mathscr{F}_t)_{t\in[0,\tau]}$ satisfait les conditions habituelles : croissante, continue à droite et complète (Liptser and Shiryayev (1989)). On note par $N = N(t)_{t\in[0,\tau]}$ un processus de comptage qui est un processus à trajectoires $p.s.$ croissantes par saut, d'amplitude 1, continues à droites et nulles à l'instant 0. Notons que $N(t+dt) - N(t)$ représente le nombre d'événements qui se produisent dans l'intervalle $[t, t+dt)$ ce qui explique bien la notion de comptage pour ce type de processus. L'intensité du processus de comptage N est définie par

$$\psi(t) = \lim_{dt\downarrow 0} \frac{\mathbb{P}[N(t+dt) - N(t) = 1 | \mathscr{F}_{t\_}]}{dt}.$$

Puisque $N(t)$ est une fonction croissante $\mathscr{F}_t$-adaptée, elle définit une sous-martingale par rapport à la filtration $\mathscr{F}_t$ (Liptser and Shiryayev (1989)). La décomposition de Doob-Meyer (Meyer) indique que tout processus de comptage peut se décomposer de façon unique comme la somme d'une martingale $M(t)$ et d'un compensateur $\Psi(t)$ tel que

$$N(t) = \Psi(t) + M(t).$$

Le compensateur $\Psi(t)$ est appelé processus d'intensité cumulée, il est croissant et prévisible, c'est-à-dire que sa valeur est connue juste avant l'instant t, et vérifie

$$\Psi(t) = \mathbb{E}[N(t)|\mathscr{F}_{t_-}] = \int_0^t \psi(s)ds.$$

En présence d'un vecteur X de covariable dans $\mathbb{R}^p$ supposé $\mathscr{F}_0$-mesurable, on définit la filtration engendrée par N et X, $\mathscr{F}_t = \sigma\{N(s), X : 0 \leq s \leq t\}$. Dans ce cas, l'intensité sera notée $\psi(t,X)$, et nous considérons :

— le modèle d'intensité multiplicative d'Aalen (Aalen (1978)) défini par

$$\psi(t,X) = \lambda(t,X)Y(t), \tag{1.8}$$

— le modèle de Cox (Cox (1972)) défini par

$$\psi(t,X) = \exp(\lambda(t,X))Y(t), \tag{1.9}$$

où λ est une une fonction positive, déterministe et bornée, et $Y(t)$ un processus stochastique positif, prévisible et borné. En pratique, on dispose de n réalisations $(X_1,N_1),\ldots,(X_n,N_n)$: si $\lambda(t,X_i) = \ldots = \lambda(t,X_n) = \lambda_0(t)$, il est clair que $N_.(t) = \sum_{i=1}^n N_i(t)$ est un processus de comptage d'intensité $\lambda_0(t)Y_.(t)$ avec $Y_.(t) = \sum_{i=1}^n Y_i(t)$.

Ces modèle incluent plusieurs exemples importants en pratiques : données censurées, processus de Poisson marqué, processus de Markov, voir Karr (1991) et Andersen et al. (1993) pour une liste détaillée. On se contente de rappeler ici le cas des données censurées : soient $T_1,\ldots,T_n$ et $C_1,\ldots,C_n$ des durées de survie et des durées de censure des n individus considérés. On les consiste en n couples de variables (Z_i,δ_i,X_i) telles que $Z_i = T_i \wedge C_i$ représente la date de l'événement terminal pour le i-ème individu, et $\delta_i = \mathbb{1}(T_i \leq C_i)$ est l'indicateur de la censure. Les processus à considérer sont donnés par $N_i(t) = \mathbb{1}(Z_i \leq t, \delta_i = 1)$ et $Y_i(t) = \mathbb{1}(Z_i \geq t)$, pour tout $i = 1,\ldots,n$. L'ensemble d'apprentissage est $\mathscr{D}_n = \{(N_i(t), Y_i(t), X_i) : i = 1,\ldots,n\}$. Une hypothèse classique habituellement considérée consiste à supposer la durée de survie T indépendante de la durée de censure C conditionnellement au vecteur de covariable X (voir Stute (1996) et Heuchenne and Van Keilegom (2007)). Le triplet (T,C,X) est une copie de $(T_i,C_i,X_i)_{1\leq i\leq n}$ de fonctions de répartition respectives F_T, G et F_X. Sous cette hypothèse, l'intensité $\lambda(t,x)$ est égale au risque instantané conditionnel de T sachant $X = x$, donné par $\lambda(t,x) = f_{T|X}(t,x)/(1-F_{T|X}(t,x))$ où $f_{T|X}$ est la densité et $F_{T|X}$ est la fonction de répartition conditionnelle de T sachant X.

1.2.2 Estimation de l'intensité d'un processus de comptage

Dans cette sous-section, nous présentons un état de l'art pour les méthodes d'estimation de la fonction d'intensité dans les modèles (1.8) et (1.9). L'estimation non-paramétrique de l'intensité cumulée, en présence de covariables, $\int_0^t \lambda(s,x)ds$ a été initié par Beran (1981). Des extensions de ces résultats ont été étudiées dans Dabrowska (1987), McKeague and Utikal (1990), et Li and Doss (1995). L'estimation semi-paramétrique de $\lambda(t,x)$ a débuté avec Cox (1972). D'autres développements ont été proposés par Huang (1999) et Linton et al. (2003). En absence de covariables, la méthode de validation croisée a été suggérée par Grégoire (1993) pour choisir la taille de fenêtre d'un estimateur à noyaux proposé par Ramlau-Hansen (1983). Patil and Wood (2004) ont élaboré une approche basée sur les

ondelettes pour estimer l'intensité (1.8). D'autres procédures dites adaptatives permettent de construite des estimateurs qui s'adaptent à la régularité de l'intensité. Par exemple, la sélection de modèle étudiée dans Reynaud-Bouret (2003, 2006), Birgé (2007), et Baraud and Birgé (2009) pour l'estimation non-paramétrique de l'intensité d'un processus de Poisson (sur des espace généraux), et par seuillage dans des bases d'ondelettes par Reynaud-Bouret and Rivoirard (2008). Comte et al. (2011) ont aussi considéré la sélection de modèles pour estimer une fonction de risque non-paramétrique avec covariables. L'estimation adaptative de la densité conditionnelle pour des données censurées a été considérée dans Antoniadis et al. (1999), Brunel and Comte (2005), et Brunel et al. (2007). Pour l'estimation paramétrique en grande dimension de (1.8), Martinussen and Scheike (2009) considèrent la sélection de covariables, Gaïffas and Guilloux (2012) proposent une procédure Lasso avec poids dépendant éventuellement des observations ("data driven") pour estimer (1.8). Lemler (2013) et Huang et al. (2013) ont étudié le Lasso pour estimer (1.9).

1.3 Contributions

Nous décrivons ici les contributions de ce manuscrit qui résument les travaux présentés dans les Chapitres 2 à 4.

1.3.1 Chapitre 2 : Apprentissage pour l'intensité d'événements récurrents avec points de rupture

Nous considérons le problème d'estimation de l'intensité $\lambda_0(t)$ qui est fonction positive, continue à droite et admettant une limite à gauche (càdlàg), d'un processus de comptage $\{N(t), t \in [0,1]\}$ à partir d'un $n-$échantillon, $N_1, \dots, N_n$ de N. Nous travaillons sous une hypothèse a priori de segmentation sparse. Autrement dit, l'intensité $\lambda_0(t)$ peut être approximée par une fonction constante par morceaux, i.e.,

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbb{1}_{J_\ell}(t),$$

pour tout $0 \le t \le 1$. Les coefficients $L_0 \ge 1$ et $\beta_{0,\ell}$ sont positifs, et $J_\ell = (\tau_{0,\ell-1}, \tau_{0,\ell}]$ pour tout $\ell = 1, \dots, L_0$, $\tau_{0,0} = 0 < \tau_{0,1} < \cdots < \tau_{0,L_0-1} < \tau_{0,L_0} = 1$.

Notre approche consiste à formuler la détection de ruptures comme un problème de sélection de variables. Pour ce faire, nous considérons des estimateurs basés sur la minimisation d'un risque empirique naturellement associé à ce modèle en ajoutant une pénalisation par variation totale avec poids. Nous fixons $m = m_n \ge 1$ un entier dépendant de n, et nous définissons Λ_m l'ensemble des fonctions constantes par morceaux sur l'intervalle $[0,1]$ par

$$\Lambda_m = \left\{ \lambda_\beta = \sum_{j=1}^{m} \beta_{j,m} \lambda_{j,m} : \beta = [\beta_{j,m}]_{1 \le j \le m} \in \mathbb{R}_+^m \right\} \text{ avec } \lambda_{j,m} = \sqrt{m} \mathbb{1}_{I_{j,m}} \text{ et } \quad I_{j,m} = \left(\frac{j-1}{m}, \frac{j}{m} \right].$$

Nous utilisons le contraste des moindres carrées

$$R_n(\lambda_\beta) = \int_0^1 \lambda_\beta(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \int_0^1 \lambda_\beta(t) dN_i(t),$$

qui est un critère d'ajustement classique dans ce modèle, voir Reynaud-Bouret (2003, 2006) et Gaïffas and Guilloux (2012). Nous introduisons la pénalisation par variation totale avec poids

$$\|\beta\|_{\mathrm{TV},\hat{w}} = \sum_{j=2}^{m} \hat{w}_j |\beta_j - \beta_{j-1}|,$$

pour tout $\beta = (\beta_j)_{1\leq j\leq m} \in \mathbb{R}^m$. Les poids $\hat{w}_j$ sont positifs et dépendent éventuellement des données tels que $\hat{w}_1 = 0$ et

$$\hat{w}_j \approx \sqrt{\frac{m\log m}{n}\bar{N}_n\Big(\Big(\frac{j-1}{m},1\Big]\Big)},$$

pour $j = 2,\dots,p$, avec $\bar{N}_n = 1/n\sum_{i=1}^n N_i$. Ils permettent de contrôler la sparsité des différences successives de β et fournissent une calibration fine de la relaxation convexe de l'hypothèse a priori de segmentation sparse. L'estimateur est donné par $\hat{\lambda} = \lambda_{\hat{\beta}}$ où

$$\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}^m_+} \{R_n(\lambda_\beta) + \|\beta\|_{\mathrm{TV},\hat{w}}\}. \tag{1.10}$$

Résultats théoriques. Le Théorème 2.3.1 dans le Chapitre 2 fournit une inégalité oracle à vitesse lente d'ordre $1/\sqrt{n}$, et vérifiée en toute généralité (sans aucune hypothèse sur l'intensité λ_0). Sous l'hypothèse que le nombre des points de rupture estimés $\hat{L} \leq L_{\max}$, nous obtenons une inégalité oracle à vitesse rapide d'ordre de $1/n$, voir Théorème 2.3.3. Un corollaire du Théorème 2.3.1 montre une borne supérieure du risque quadratique de (1.10) est de l'ordre

$$\frac{\Delta^2_{\beta,\max}}{m} + \frac{m\log m}{n},$$

où $\Delta_{\beta,\max}$ représente la taille maximale des sauts de l'intensité λ_0. Par conséquent, un bon choix pour m est de l'ordre de $\sqrt{n}$.

Procédure de la détection des points de rupture. La consistance en détection nécessite le fait que les points de rupture soient séparés au moins d'une distance égale à $1/m$ qui est l'ordre de la haute résolution dans ce problème. Relativement à la taille de la grille de segmentation, nous définissons une suite de *points de rupture approximés* $[j_\ell]_{0\leq\ell\leq L_0}$ comme étant la borne droite de l'unique intervalle $I_{j_\ell,m}$ contenant le vrai point de rupture $\tau_{0,\ell}$, i.e., $\tau_{0,\ell} \in \left(\frac{j_\ell-1}{m}, \frac{j_\ell}{m}\right]$ pour tout $\ell = 1,\dots,L_0-1$. Nous définissons $\hat{\tau}_\ell = \hat{j}_\ell/m$ pour tout $\ell = 0,\dots,\hat{L}+1$.

Notre premier résultat de consistance est établi sous l'hypothèse que le nombre des points de rupture estimés soit le bon, i.e., $\hat{L} = L_0 - 1$. Le Théorème 2.4.4 montre que si $m = \sqrt{n}, \varepsilon_n = 1/\sqrt{n}$ et $\Delta_{\beta,\min} = n^{-1/6}$, alors les points de rupture estimés $\{\hat{\tau}_1,\dots,\hat{\tau}_{\hat{L}}\}$ satisfont

$$\mathbb{P}\Big[\max_{1\leq\ell\leq L_0-1} |\tau_{0,\ell} - \hat{\tau}_\ell| \leq \varepsilon_n\Big] \to 1, \text{ quand } n \to +\infty.$$

Notre deuxième résultat consiste à raffiner l'hypothèse $\hat{L} = L_0 - 1$. Dans le Théorème 2.4.5, nous prouvons que notre procédure est consistante même si le nombre des points de rupture est sur-estimé. Pour cela, nous évaluons la distance non-symétrique de Haussdorf entre les ensembles des vrais et estimés points de rupture.

Nous soulignons que le problème de détection dans ce travail diffère de celui du cas standard signal + bruit gaussien (cf. Harchaoui and Lévy-Leduc (2010)). En effet, nous

cherchons à détecter les points de rupture dans la fonction d'intensité en temps continu et en présence d'un inévitable biais d'approximation.

Résultats pratiques. Le problème (1.10) s'écrit sous la forme

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}_+^m} \left\{ \frac{1}{2} \|\mathbf{N} - \beta\|_2^2 + \|\beta\|_{\mathrm{TV},\hat{w}} \right\}, \tag{1.11}$$

avec $\mathbf{N}$ est le vecteur de coordonnées $\mathbf{N}_j = \bar{N}_n(I_{j,m})$ pour tout $j = 1, \ldots, m$. Ceci est équivalent à $\hat{\beta} = \operatorname{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}(\mathbf{N})$. Dans l'Algorithme 3 du Chapitre 2, nous considérons une extension de l'approche utilisée par Condat (2013) pour calculer d'une manière directe l'opérateur proximal de la variation totale avec poids. Nous construisons cet algorithme en se basant sur le problème dual de (1.11), et les conditions d'optimalité de Karush-Kuhn-Tucker (Boyd and Vandenberghe (2004)).

Nous illustrons notre procédure de détection sur un jeu de données des cellules cancéreuses (cancer de seins) HCC1954 et saines BL1954. Elles sont produites, explorées par Chiang et al. (2009) et elles contiennent 7.72 millions de reads pour HCC1954 et 6.65 millions pour BL1954 dont chaque read est de longues 36 paire de base, voir Figures 2.4 et 2.5 dans le Chapitre 2. Nous identifions ici l'axe du temps à la position dans le génome. Dans la Figure 2.6 du Chapitre 2, nous traçons la meilleur solution en utilisant les procédures de variation totale avec et sans poids. Nous constatons que celle avec poids a de bonnes performances : l'intensité est lisse et les positions des points de rupture sont plus visibles. Une caractéristique importante de notre approche est le temps d'exécution de l'algorithme proposé. En effet, la solution est obtenue en moins de quelques millisecondes sur un ordinateur portable moderne.

1.3.2 Chapitre 3 : Binarsity : prédiction en grande dimension via la sparsité induite par la binarisation de variables

En apprentissage supervisé, un pre-processing sur les variables explicatives est souvent nécessaire, en particulier pour les méthodes de type modèles linéaires généralisés. La technique habituelle consiste à standardiser et réduire les colonnes de la matrice $\boldsymbol{X}$ de variables explicatives. Une autre approche très utilisée en pratique (Dougherty et al. (1995)) consiste à discrétiser ces variables. Dans le Chapitre 3, nous proposons un pre-processing de $\boldsymbol{X}$ à l'aide d'un processus de binarisation, et nous considérons le problème de prédiction en grande dimension via sur une approche par pénalisation induite par la binarisation.

Binarisation de variables. On utilise la notation suivante : $\boldsymbol{X}_{\bullet,j}$ la j-ème variable en colonne et $\boldsymbol{X}_{i,\bullet}$ la i-ème variable en ligne de $\boldsymbol{X}$. La matrice binarisée de $\boldsymbol{X}$ notée $\boldsymbol{X}^B$ est la matrice à d colonnes ($d \gg p$) où on remplace $\boldsymbol{X}_{\bullet,j}$ par d_j colonnes $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j}$ ne contenant que des 0 ou 1. Si la colonne $\boldsymbol{X}_{\bullet,j}$ prend des valeurs discrètes dans un ensemble de modalités $\{1, \ldots, M_j\}$, alors on pose $d_j = M_j$ et

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1, & \text{if } \boldsymbol{X}_{i,j} = k, \\ 0, & \text{sinon,} \end{cases}$$

pour $i = 1,\dots,n$ et $k = 1,\dots,d_j$. Si la colonne $\boldsymbol{X}_{\bullet,j}$ est quantitative, alors on considère une partition d'intervalles formés par des quantiles de cette dernière, $I_{j,1},\dots,I_{j,d_j}$, tels que pour tout $k = 1,\dots,d_j$, $I_{j,k} = \left[q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right)$ avec $q_j(\alpha)$ le quantile d'ordre α de $\boldsymbol{X}_{\bullet,j}$ et on pose

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1, & \text{si } \boldsymbol{X}_{i,j} \in I_{j,k}, \\ 0, & \text{sinon.} \end{cases}$$

Ce passage de la matrice $\boldsymbol{X}$ à la matrice $\boldsymbol{X}^B$ s'appelle binarisation. C'est une technique très utilisée dans de nombreux domaines notamment en marketing digital (Chapelle et al. (2014)). L'idée est qu'en "éclatant" une variable en plusieurs variables binaires, nous obtenons une meilleure attache au données par une réponse non-linéaire par rapport aux variables d'origine. Dans le Chapitre 3, nous introduisons alors une pénalisation forçant les coefficients associés à une variable binarisée à ne pas prendre un très grand nombre de valeurs différentes.

A chaque variable binarisée $\boldsymbol{X}^B_{\bullet,j,k}$ correspond un coefficient $\theta_{j,k}$. Le paramètre de la binarisation de la j-ème variable est un vecteur noté $\theta_{j,\bullet} = [\theta_{j,1} \cdots \theta_{j,d_j}]^\top$. Nous considérons la concaténation de ces vecteurs en un seul $\theta = [\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top]^\top$ de taille $d = \sum_{j=1}^p d_j$. Une illustration de vecteur θ est donnée dans la Figure 1.2.

Remarque importante. La matrice binarisée $\boldsymbol{X}^B$ n'est pas de rang plein, puisque dans chaque bloc la somme des colonnes $\boldsymbol{X}^B_{\bullet,j,1},\dots,\boldsymbol{X}^B_{\bullet,j,d_j}$ est égale à $\mathbf{1}_n$, la colonne ne contenant que des 1 (intercept). Pour remédier à cette sur-paramétrisation, il nous faut rajouter des contraintes. Ces contraintes sont directement liées à l'interprétation des prédicteurs. Par exemple, on peut supposer que $\theta_{j,k} = 0$ pour un certain $k \in \{1,\dots,d_j\}$ ou bien $\sum_{k=1}^{d_j} \theta_{j,k} = 0$ (voir Agresti (2015)). Dans notre cas, la notion de sparsité que l'on cherche à induire est alors la suivante : tout bloc $\theta_{j,\bullet}$ peut être constant ou contient un nombre assez petit de valeurs différentes. Nous choisissons donc de travailler sous la contrainte $\sum_{k=1}^{d_j} \theta_{j,k} = 0$.

La nouvelle notion de sparsité, nommée *binarsity*, est donnée par

$$\begin{aligned} \text{bina}(\theta) &= \sum_{j=1}^p \left(\|\theta_{j,\bullet}\|_{\text{TV}} + \delta_{\mathscr{H}_j}(\theta_{j,\bullet}) \right) \\ &= \sum_{j=1}^p \Big(\sum_{k=2}^{d_j} |\theta_{j,k} - \theta_{j,k-1}| + \delta_{\mathscr{H}_j}(\theta_{j,\bullet}) \Big), \end{aligned} \tag{1.12}$$

où $\mathscr{H}_j = \{\beta_{j,\bullet} \in \mathbb{R}^{d_j} : \sum_{k=1}^{d_j} \beta_{j,k} = 0\}$ est l'hyperplan de $\mathbb{R}^{d_j}$ de vecteur normale $\mathbb{1}_{j,\bullet} = (1,\dots,1)^\top$, et la fonction indicatrice est donnée par

$$\delta_{\mathscr{H}_j}(\beta_{j,\bullet}) = \begin{cases} 0, & \text{si } \beta_{j,\bullet} \in \mathscr{H}_j, \\ \infty, & \text{sinon.} \end{cases}$$

Si la j-ème variable est statistiquement non pertinente pour la prédiction, alors le bloc $\theta_{j,\bullet}$ qui lui correspond est constant, et dans ce cas sa contribution à binarsity est égale à zéro. Si la j-ème variable est pertinente alors le nombre des valeurs différentes dans le bloc $\theta_{j,\bullet}$ doit être assez petit pour un bon compris biais-variance.

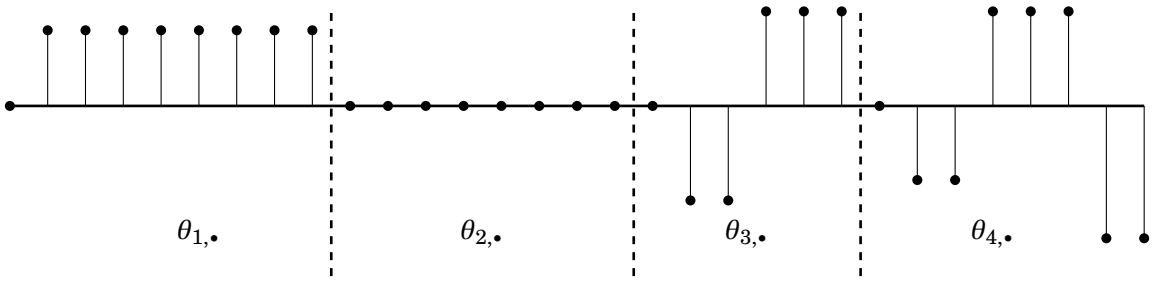


Fig. 1.2 – Illustration de $\theta = [\theta_{1,\bullet}^\top, \ldots, \theta_{p,\bullet}^\top]^\top$ avec : $p = 4, d_1 = 9$, $d_2 = 8$, $d_3 = 6$, $d_4 = 8$.

Apprentissage sous l'hypothèse de binarsity. Nous considérons une collection $\mathscr{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ de copies $i.i.d.$ d'une variable (X, Y) à valeurs dans $\mathbb{R}^p \times \mathscr{Y}$ d'un modèle linéaire généralisé où la distribution de Y conditionnellement à $X = x$ appartient à une famille exponentielle (Nelder and Wedderburn (1972)), i.e.,

$$f(y; m_0(x)) = \exp\big(y m_0(x) - b(m_0(x))\big).$$

La fonction $b(\cdot)$ est supposée connue, tandis que la fonction $m_0(\cdot)$ ne l'est pas. Pour estimer m_0, nous construisons des estimateurs dans l'espace engendré par les variables binarisées en minimisant un risque empirique pénalisé. Plus précisément, nous considérons la log-vraisemblance négative du modèle linéaire généralisé

$$R_n(m_\theta) = R_n(\theta) = \frac{1}{n}\sum_{i=1}^n -\boldsymbol{Y}_i\, m_\theta(\boldsymbol{X}_{i,\bullet}) + b(m_\theta(\boldsymbol{X}^B_{i,\bullet})),$$

avec $m_\theta(\boldsymbol{X}_{i,\bullet}) = \langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle$. Nous introduisons une version pondérée de bina$\cdot$ appelée bina$_{\hat{w}}\cdot$, en ajoutant des poids $\hat{w}_{j,\bullet}$ dans chaque bloc $\theta_{j,\bullet}$. Elle est donnée par

$$\begin{aligned}\text{bina}_{\hat{w}}(\theta) &= \sum_{j=1}^p \Big(\|\theta_{j,\bullet}\|_{\text{TV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}(\theta_{j,\bullet})\Big) \\ &= \sum_{j=1}^p \Big(\sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| + \delta_{\mathscr{H}_j}(\theta_{j,\bullet})\Big).\end{aligned}$$

Les poids permettent de contrôler la sparsité des différences successives dans chaque bloc $\theta_{j,\bullet}$, et fournissent une calibration fine de binarsity. Ils vérifient pour tout $j = 1, \ldots, p$, $\hat{w}_{j,1} = 0$ et pour tout $k \in \{2, \ldots, d_j\}$ ils sont de l'ordre de $\hat{w}_{j,k} \approx \sqrt{\frac{d_{\max}\hat{n}_{j,k}}{n}}$ avec $d_{\max} = \max_{j=1,\ldots,p} d_j$, et

$$\hat{n}_{j,k} = \frac{\#\Big(\Big\{i = 1, \ldots, n : \boldsymbol{X}_{i,j} \in \big[q_j\big(\tfrac{k}{d_j}\big), q_j(1)\big]\Big\}\Big)}{n}.$$

Nous définissons $\hat{m} = m_{\hat{\theta}}$ avec

$$\hat{\theta} = \text{argmin}_{\theta \in \mathbb{R}^d} \big\{R_n(\theta) + \text{bina}_{\hat{w}}\,\theta\big\}. \tag{1.13}$$

Pour évaluer la qualité de l'estimation, nous utilisons le risque d'excès $R(m_{\hat{\theta}}) - R(m_0) = \mathbb{E}[R_n(m_{\hat{\theta}})] - \mathbb{E}[R_n(m_0)]$. Nous remarquons que ce dernier se réécrit sous la forme d'une distance de Kullback empirique $KL_n(m_0, m_{\hat{\theta}})$, donnée par

$$KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) = \frac{1}{n}\sum_{i=1}^n KL\big[f(y; m_0(x_{i,\bullet})), f(y; m_{\hat{\theta}}(x_{i,\bullet}))\big].$$

Résultats. Dans le Chapitre 3, la Proposition 3.3.6 est une inégalité oracle à vitesse lente vérifiée par l'estimateur (1.13). Sous une hypothèse de valeurs propres restreintes sur la matrice binarisée $\boldsymbol{X}^B$, nous prouvons une inégalité oracle non exacte à vitesse rapide d'ordre $\log d/n$ (voir Théorème 3.3.7). L'Algorithme 4 dans le le Chapitre 3 calcule l'opérateur proximal de la pénalité induite par binarsity, $\mathrm{prox}_{\mathrm{bina}}$.. Par conséquent, l'estimateur (1.13) peut être calculé à l'aide de la procédure FISTA.

1.3.3 Chapitre 4 : Modèles d'Aalen et de Cox en grande dimension avec des covariables temps-dépendantes

Nous observons n copies indépendantes $\{N_i(t), Y_i(t), X_i(t) : i = 1, \dots, n, 0 \le t \le \tau\}$. Le vecteur de covariables $X_i(t) = (X_i^1(t), \dots, X_i^p(t)) \in \mathbb{R}^p$ dépend du temps, $N_i(t)$ un processus de comptage, $Y_i(t)$ un processus aléatoire à valeurs dans $[0,1]$ et $[0,\tau]$ un intervalle d'étude avec τ le temps terminal. À partir de ces observations, nous cherchons à estimer la fonction du risque dans le les modèles suivants :

— Modèle d'Aalen

$$\lambda_\star^{\mathrm{A}}(t, X(t)) = X(t)\beta^\star(t),$$

— Modèle de Cox

$$\lambda_\star^{\mathrm{M}}(t, X(t)) = \exp\big(X(t)\beta^\star(t)\big),$$

avec $\beta^\star$ est une fonction à p variables de $[0,\tau]$ à valeurs dans $\mathbb{R}^p$ que l'on cherche à estimer. Pour cela, nous considérons des estimateurs basés sur des histogrammes, voir Murphy and Sen (1991). Plus précisément, nous disposons d'une L-partition ($L \in \mathbb{N}^*$) de l'intervalle de temps $[0,\tau]$ définie par

$$\varphi_l = \sqrt{\frac{L}{\tau}} \mathbb{1}(I_l) \text{ avec } I_l = \Big(\frac{l-1}{L}\tau, \frac{l}{L}\tau\Big].$$

Pour tout $j = 1, \dots, p$, un candidat pour estimer le j-ème coefficient $\beta_j^\star$ de $\beta^\star$ appartient à l'ensemble de fonctions constantes par morceaux

$$\mathscr{H}_L = \Big\{\alpha(\cdot) = \sum_{l=1}^{L} \alpha_l \varphi_l(\cdot) : (\alpha_l)_{1 \le l \le L} \in \mathbb{R}_+^L\Big\}.$$

Pour chaque individu i nous lui associons un processus de covariables p-dimensionnel $X_i(t)$, et nous notons par $X_i^j(t)$ le processus associé pour son j-ème covariable. Pour toute fonction à p variables β, estimateur candidat de $\beta^\star$, nous désignons par β_j sa j-ème variable (fonction). Nous définissons l'ensemble d'estimateurs candidats par

$$\Lambda^{\mathrm{A}} = \{x, t \in [0,\tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = x(t)\beta(t) \mid \forall j\ \beta_j \in \mathscr{H}_L\}$$

pour le modèle d'Aalen et par

$$\Lambda^{\mathrm{M}} = \{x, t \in [0,\tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = \exp\big(x(t)\beta(t)\big) \mid \forall j\ \beta_j \in \mathscr{H}_L\}$$

pour le modèle de Cox. Pour l'ensemble Λ^{M} ou Λ^{A}, chaque coefficient est une fonction constante par morceaux. β est considérée à la fois comme une fonction à p variables ou comme un vecteur de dimension $p \times L$ défini par

$$\beta = (\beta_{1,\cdot}^\top, \dots, \beta_{p,\cdot}^\top)^\top = (\beta_{1,1}, \dots, \beta_{1,L}, \dots, \beta_{p,1}, \dots, \beta_{p,L})^\top,$$

où $\beta_{j,\cdot}$ appartient à $\mathbb{R}^L$ et $\beta_{j,l}$ est la valeur prise par la j-ème coordonnée dans le l-ème intervalle de notre L-partition $\{I_1,\dots,I_L\}$. Nous considérons la minimisation des fonctionnelles suivantes : les moindres carrées pour le modèle d'Aalen

$$\ell_n^{\mathrm{A}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \big(\lambda_\beta^{\mathrm{A}}(t,X_i(t))\big)^2 Y_i(t)dt - 2\int_0^\tau \lambda_\beta^{\mathrm{A}}(t,X_i(t))dN_i(t)\Big\},$$

et la log-vraisemblance pour le modèle de Cox

$$\ell_n^{\mathrm{M}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\big(\lambda_\beta^{\mathrm{M}}(t,X_i(t))\big)dN_i(t) - \int_0^\tau Y_i(t)\lambda_\beta^{\mathrm{M}}(t,X_i(t))dt\Big\}.$$

Nous introduisons une pénalité $(\ell_1+\ell_1)$-variation totale avec poids défini par

$$\|\beta\|_{\mathrm{gTV},\hat\gamma} = \sum_{j=1}^{p}\big(\hat\gamma_{j,1}|\beta_{j,1}| + \sum_{l=2}^{L}\hat\gamma_{j,l}|\beta_{j,l}-\beta_{j,l-1}|\big)$$

pour tout $\beta\in\mathbb{R}^{p\times L}$ avec $\hat\gamma = (\hat\gamma_{1,\cdot}^\top,\dots,\hat\gamma_{p,\cdot}^\top)^\top$, tel que $\hat\gamma_{j,\cdot}\in\mathbb{R}_+^L$ pour tout $j=1,\dots,p$, donné par

$$\hat\gamma_{j,l} \approx \sqrt{\frac{L\log(pL)}{n}\hat V_{j,l}}, \text{ avec } \hat V_{j,l} = \frac{1}{n}\sum_{i=1}^{n}\int_{\bigcup_{u=l}^{L} I_u}(X_i^j(t))^2 dN_i(t).$$

Nos estimateurs sont définis par $\hat\lambda^{\mathrm{A}} = \lambda^{\mathrm{A}}_{\hat\beta^{\mathrm{A}}}$ et $\hat\lambda^{\mathrm{M}} = \lambda^{\mathrm{M}}_{\hat\beta^{\mathrm{M}}}$ où

$$\hat\beta^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}}\Big\{\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat\gamma}\Big\}, \tag{1.14}$$

et

$$\hat\beta^{\mathrm{M}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}}\Big\{\ell_n^{\mathrm{M}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat\gamma}\Big\}. \tag{1.15}$$

Résultats

Les Théorèmes 4.3.1 et 4.7.3 sont des inégalités oracles à vitesse lente vérifiées par $\hat\lambda^{\mathrm{A}}$ et $\hat\lambda^{\mathrm{M}}$ avec une grande probabilité. La résolution algorithmique des problèmes (1.14) et (1.15) s'est fait par l'implémentation d'un algorithme proximal de descente du gradient stochastique, voir Algorithme 6 dans le Chapitre 4. Nous illustrons nos méthodes sur des données simulées et réelles en les comparons avec les procédures étudiées dans Martinussen and Scheike (2007).

Chapter 2

Learning the Intensity of Time Events with Change-Points

This chapter is an extended version of the article Alaya et al. (2015) published in IEEE Transactions on Information Theory.

Abstract

We consider the problem of learning the inhomogeneous intensity of a counting process, under a sparse segmentation assumption. We introduce a weighted total-variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. We prove that this leads to a sharp tuning of the convex relaxation of the segmentation prior, by stating oracle inequalities with fast rates of convergence, and consistency for change-points detection. This provides first theoretical guarantees for segmentation with a convex proxy beyond the standard i.i.d signal + white noise setting. We introduce a fast algorithm to solve this convex problem. Numerical experiments illustrate our approach on simulated and on a high-frequency genomics dataset.

Contents

2.1 Introduction

Counting processes are widely used in engineering to describe systems where stochastic events occur, such as genomics, biology, econometrics, communications and networks, see Andersen et al. (1993). In these problems, the aim is to estimate the intensity function, which determines the instantaneous rate of occurrence of an event. In the statistical literature, this topic has been extensively discussed in several previous works. Procedures based on kernel estimation Ramlau-Hansen (1983), cross-validation Grégoire (1993), wavelet methods Patil and Wood (2004), local polynomial estimators Chen et al. (2011), model selection Reynaud-Bouret (2003), etc. are considered for the non-parametric estimation of the intensity.

In this paper, we want to recover the intensity $\lambda_0(t)$ of a counting process $\{N(t), t \in [0,1]\}$ from n observations of N. We work under the assumption that λ_0 can be well-approximated by a piecewise constant function, and we deal with this problem with a signal segmentation point-of-view, where the goal is to find the unknown times of abrupt changes in the dynamic of the signal. This is referred to *multiple change-point problem* in statistical literature, see Khodadadi and Asgharian (2008) for a recent review with interesting references. A change-point is a time or position where the structure of the object changes and the goal of change-point detection is to estimate these positions.

Several examples of practical importance fulfill the model of multiple change-points. A particularly interesting example comes from the next-generation sequencing (NGS) DNA process. Indeed, an important application of NGS technologies is the study of the transcriptome and the resulting experiment is called RNA-seq. In a typical RNA-seq experiment, a sample of RNA is amplified, shattered, and converted to a library of a cDNA fragments. Then, it is sequenced on a high-throughput platform which is available commercially. Finally, the raw data result in large amounts of DNA fragments sequences called reads. These reads are then mapped to the reference genome by an appropriate algorithm, that tells us the region from which each

read comes from. RNA-seq can be modeled mathematically as replications of an inhomogeneous counting process with a piecewise constant intensity Shen and Zhang (2012). The counting process counts the number of reads whose first base maps to the left base of a given chromosome's location. In Shen and Zhang (2012), a Bayesian approach for the detection of change-points is considered. Other approaches based on Bayesian model-based clustering and segmentation are given in Picard et al. (2007).

In the present paper, we consider the estimation of $\tau_{0,\ell}$ and $\beta_{0,\ell}$ in the following model:

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbb{1}_{(\tau_{0,\ell-1},\tau_{0,\ell}]}(t) \tag{2.1}$$

for $0 \leq t \leq 1$, with the convention $\tau_{0,0} = 0$ and $\tau_{0,L_0} = 1$. Our approach consists in reframing this task as a variable selection task. We introduce a penalized least-squares criterion with a data-driven total-variation penalization, which is ℓ_1-penalization of the discrete gradient of the parameter.

This convex proxy for segmentation with an extra ℓ_1-penalization for sparsity, called *fused Lasso*, is introduced in Tibshirani et al. (2005). Theoretical guarantees for this procedure are given in Harchaoui and Lévy-Leduc (2010) in the white noise setting, for the segmentation of a one-dimensional signal. A group fused Lasso is introduced in Bleakley and Vert for the detection of multiple change-points shared by a set of co-occurring one-dimensional signals, and an algorithm is derived to solve the corresponding convex problem. The determination of the number of structural changes in multitask learning via the group fused Lasso is considered in Qian and Su. (2013).

Beyond the one-dimensional setting, total-variation penalization is well-known and commonly used in image denoising, deblurring and segmentation, see for instance Chambolle and Darbon (2009) and Chambolle et al. (2010). In this context, one needs to define a graph of neighboring nodes (pixels), and the problem can be solved efficiently by reformulating it as a min-cut problem and solving it using a max-flow algorithm Hochbaum (2001).

Other close references are the following: Gaïffas and Guilloux (2012) proves sharp oracle inequalities for the Lasso in hazards models, Ciupera studies Lasso-type estimators in a linear regression model with multiple change-points, Rinaldo (2009) considers denoising of a sparse and block signal, Boysen et al. (2009) studies the asymptotics for jump-penalized least squares regression aiming at approximating a regression function by piecewise constant functions. An algorithm of majorization-minimization for high dimensional fused Lasso regression is proposed in Yu et al. (2015), a testing approach for the segmentation of the hazard function is given in Goodman et al. (2011).

The papers Reynaud-Bouret (2003), Tibshirani et al. (2005), Harchaoui and Lévy-Leduc (2010), Bleakley and Vert, Qian and Su. (2013), are most relevant to our work. In Reynaud-Bouret (2003), a model selection procedure is introduced to estimate the intensity function. In Harchaoui and Lévy-Leduc (2010) and Tibshirani et al. (2005), the authors propose an adaptation of the Lasso algorithm to detect change-points in the standard i.i.d signal + Gaussian white noise framework. In Bleakley and Vert

and Qian and Su. (2013), the authors use group fused Lasso to solve the structural change-points in linear regression problems. This paper is different from these works in the following aspects. First, a main feature of our results is that they are derived for a signal in continuous time, as compared to Harchaoui and Lévy-Leduc (2010), Qian and Su. (2013) and Tibshirani et al. (2005). Namely, we aim at detecting change-points in the intensity function. Hence, this problem is prone to an unavoidable non-parametric bias of approximation by a piecewise constant function, which makes our mathematical analysis very different. A second main feature of our results is that we introduce a weighted total-variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. This is not necessary in the Gaussian and discrete signal + noise setting from Harchaoui and Lévy-Leduc (2010) for instance. As a side product, we are able to use the same tuning parameters both for consistency in oracle inequalities, see Theorems 2.3.1 and 2.3.3, and detection of change-points, see Theorems 2.4.4 and 2.4.5. A third main feature of our approach is that we use a convex surrogate for the sparsity of the discrete gradient of the signal, that can be solved numerically very efficiently, see Section 2.5, even for a large signal (using many bins). This is not the case for the approach described in Reynaud-Bouret (2003), which is based on ℓ_0 model-selection techniques. Furthermore, our oracle inequalities are sharp in the sense that the leading constant in front of the bias terms is equal to one.

The rest of the paper is organized as follows. In Section 2.2, we provide basic notations. Then, we present our estimation procedure. Section 2.3 develops oracle inequalities for the estimator, see Theorems 2.3.1 and 2.3.3. Section 2.4 gives results in change-points detection, see Theorems 2.4.4 and 2.4.5. Section 2.5 describes a fast algorithm to solve the convex problem studied in the paper. The proofs of the main statements are gathered in Sections 2.6 to 2.8.

2.2 Counting processes with a sparse segmentation prior

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{0\leq t\leq 1}$ a filtration satisfying the usual conditions Liptser and Shiryayev (1989): increasing, right-continuous and complete. A *counting process* is a stochastic process $\{N(t)\}_{0\leq t\leq 1}$ which is $(\mathcal{F}_t)$-adapted to the filtration, with right-continuous and piecewise constant paths almost surely (a.s.), with jump of size +1 at event times such that $N(0) = 0$ and $N(t) < \infty$ a.s. The term counting process is natural: $N(t) - N(s)$ corresponds to the number of events of a certain type occurring in the interval $(s, t]$. The Poisson process is the most common example of a counting process, where the jumps occur randomly and independently of each other on disjoint intervals, see for instance Brémaud (1981) and Karr (1991) for references on point processes and their statistical estimation.

Since N is increasing, it is a submartingale, so it follows from the Doob-Meyer decomposition theorem Aalen (1978). Namely, $N = \Lambda_0 + M$, where Λ_0 is a predictable increasing process called the compensator of N and M is a $(\mathcal{F}_t)$-martingale. We as-

sume in the following that

$$\Lambda_0(t) = \mathbb{E}[N(t)] = \int_0^t \lambda_0(s)ds \tag{2.2}$$

for $0 \le t \le 1$, where λ_0 is a non-negative right-continuous function with left-hand limits called *intensity rate* of N. Under this assumption, $M(t) = N(t) - \int_0^t \lambda_0(s)ds$ is a local square-integrable martingale with *quadratic variation* given by $\langle M \rangle(t) = \int_0^t \lambda_0(s)ds$ and *optional variation* $[M](t) = \int_0^t \lambda_0(s)dN(s)$.

2.2.1 Sparse segmentation assumption

We work under the assumption that the intensity is piecewise constant, over unknown inhomogeneous intervals of time. From now on, $\mathbb{1}_A$ stands for the indicator function of a set A. For some results in the paper, we will use

Assumption 2.2.1. *We assume that the intensity writes*

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbb{1}_{J_\ell}(t), \, 0 \le t \le 1, \tag{2.3}$$

with $L_0 \ge 1$, $\beta_{0,\ell}$ are positive coefficients, and where $J_0 = \{0\}$, $J_\ell = (\tau_{0,\ell-1}, \tau_{0,\ell}]$ for $\ell = 1, \dots, L_0$ and $\tau_{0,0} = 0 < \tau_{0,1} < \cdots < \tau_{0,L_0-1} < \tau_{0,L_0} = 1$.

Assumption 2.2.1 means that $L_0 - 1$ changes affect the value of λ_0 at unknown instants $\tau_{0,\ell}$. The number of change-points $L_0 - 1$ is unknown. In this setting, we want to recover the intensity λ_0, by jointly estimating $L_0, \tau_{0,\ell}$ and $\beta_{0,\ell}$, for $\ell = 1, \dots, L_0 - 1$. Throughout the paper, we will assume the following.

Assumption 2.2.2. *We observe n i.i.d copies of N on $[0,1]$, denoted $N_1, \dots, N_n$.*

The assumption that the process is in $[0,1]$ is for the sake of simplicity. Assumption 2.2.2 is equivalent to observing a single process N with intensity $n\lambda_0$, which is only used to have a notion of growing observations with an increasing n.

2.2.2 A procedure based on total-variation penalization

Fix $m = m_n \ge 1$, an integer that shall go to infinity as $n \to \infty$. Let us define the set of nonnegative piecewise constant functions on $[0,1]$ given by

$$\Lambda_m = \Big\{ \lambda_\beta = \sum_{j=1}^{m} \beta_{j,m} \lambda_{j,m} : \beta = [\beta_{j,m}]_{1 \le j \le m} \in \mathbb{R}_+^m \Big\}, \tag{2.4}$$

where

$$\lambda_{j,m} = \sqrt{m} \mathbb{1}_{I_{j,m}} \quad \text{and} \quad I_{j,m} = \Big(\frac{j-1}{m}, \frac{j}{m} \Big].$$

The linear space Λ_m is endowed by the norm $\|\lambda\| = (\int_0^1 \lambda^2(t)dt)^{1/2}$. We introduce the least-squares functional

$$R_n(\lambda) = \int_0^1 \lambda(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \int_0^1 \lambda(t) dN_i(t),$$

which is the goodness-of-fit criterion to be used in this setting, see among others Reynaud-Bouret (2003). Note that $\{\lambda_{j,m} : j = 1,\dots,m\}$ produces an orthonormal basis of Λ_m, it implies that

$$R_n(\lambda_\beta) = \sum_{j=1}^{m} \beta_{j,m}^2 - \frac{2\sqrt{m}}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_{j,m} N_i(I_{j,m})$$

for any $\beta \in \mathbb{R}_+^m$. Now, let us introduce the *weighted total-variation* penalization

$$\|\beta\|_{\mathrm{TV},\hat{w}} = \sum_{j=2}^{m} \hat{w}_j |\beta_j - \beta_{j-1}| \tag{2.5}$$

for $\beta = [\beta_j]_{1\le j\le m} \in \mathbb{R}^m$, where $\hat{w} = [\hat{w}_j]_{1\le j\le m}$ is a positive vector of weights (eventually depending on data) to be defined later on, with $\hat{w}_1 = 0$. The data-driven weights $\hat{w}$ will allow to design sharp tuning of the total-variation penalization. Then, given $m \ge 1$ and a weights vector $\hat{w}$, we introduce

$$\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}_+^m} \{R_n(\lambda_\beta) + \|\beta\|_{\mathrm{TV},\hat{w}}\}, \tag{2.6}$$

hence an estimator of λ_0 is given by $\hat{\lambda} = \lambda_{\hat{\beta}}$. An estimation of the change-point locations is obtained from the support of the discrete gradient of $\hat{\beta}$. Namely, define

$$\hat{S} = \{j : \hat{\beta}_{j,m} \neq \hat{\beta}_{j-1,m} \text{ for } j = 2,\dots,m\}, \tag{2.7}$$

and denote by $\hat{L} = |\hat{S}|$ the estimated number of change-points.

We denote the mean counting process $\bar{N}_n = n^{-1}\sum_{i=1}^n N_i$, and the unweighted TV penalization by $\|\beta\|_{\mathrm{TV}} = \sum_{j=2}^m |\beta_j - \beta_{j-1}|$ for $\beta \in \mathbb{R}^m$. We use also the notation $\bar{N}_n(I) = \int_I d\bar{N}_n(t)$ for any $I \subset [0,1]$.

2.3 Sharp oracle inequalities

In this section we address the statistical properties of $\hat{\lambda}$ stated in (2.6), by proving two oracle inequalities. Theorem 2.3.1 below is an oracle inequality of "slow-type" Bickel et al. (2009) that holds in full generality, while Theorem 2.3.3 is a fast oracle inequality, that holds under the assumption that the number of the estimated change-points is upper bounded by a known constant $L_{\max}$. Both oracle inequalities are sharp in the sense that the constant term in front of the oracle term $\inf_\beta \|\lambda_\beta - \lambda\|$ is equal to one.

Theorem 2.3.1. *Fix $x > 0$ and introduce the data-driven weights,*

$$\hat{w}_j = 5.66\sqrt{\frac{m(x + \log m + \hat{h}_{n,x,j})\hat{V}_j}{n}} + 9.31\frac{\sqrt{m}(x + 1 + \log m + \hat{h}_{n,x,j})}{n},$$

where $\hat{V}_j = \bar{N}_n\big(\big(\frac{j-1}{m}, 1\big]\big)$ and

$$\hat{h}_{n,x,j} = 2\log\log\Big(\frac{6en\hat{V}_j + 14e(x + \log m)}{28(x + \log m)} \vee e\Big).$$

Then, if $\hat{\lambda}$ is given by (2.6), we have

$$\|\hat{\lambda} - \lambda_0\|^2 \le \inf_{\beta\in\mathbb{R}_+^m} \Big(\|\lambda_\beta - \lambda_0\|^2 + 2\|\beta\|_{\mathrm{TV},\hat{w}}\Big) \tag{2.8}$$

with a probability larger than $1 - 12.85e^{-x}$.

The proof of Theorem 2.3.1 is postponed in Section 2.6. We define $\beta_{0,m} = [\beta_{0,j,m}]_{1\le j\le m}$ the coefficients vector of the projection of λ_0 on Λ_m and $\Delta_{\beta,\max} = \max_{1\le \ell,\ell'\le L_0} |\beta_{0,\ell} - \beta_{0,\ell'}|$, which is the maximum jump size of λ_0. Under Assumption 2.2.1, a control of the approximation term leads to the following.

Corollary 2.3.2. *Given Assumption 2.2.1, and under the same assumptions as the ones from Theorem 2.3.1, we have*

$$\|\hat{\lambda} - \lambda_0\|^2 \le \frac{2(L_0-1)\Delta^2_{\beta,\max}}{m} + 2\|\beta_{0,m}\|_{\mathrm{TV}} \max_{1\le j\le m} \hat{w}_j. \tag{2.9}$$

The proof of Corollary 2.3.2 is given in Section 2.6. Theorem 2.3.1 uses a data-driven weighting of the TV penalization, based on weights roughly given by

$$\hat{w}_j \approx \sqrt{\frac{m\log m}{n} \bar{N}_n\Big(\Big(\frac{j-1}{m}, 1\Big]\Big)}. \tag{2.10}$$

This exhibits a new scaling of the TV penalization, which is natural and of importance in this setting. The shape of this data-driven weighting comes from a Bernstein's concentration with data-driven variance, necessary for the control of the noise term (a martingale with jumps), given in Proposition 2.6.1 below, see Section 2.6.1.

Theorem 2.3.3. *Fix $x > 0$ and let $\hat{\lambda}$ be the same as in Theorem 2.3.1. Assume that the estimated number of change-points $\hat{L}$ satisfies $\hat{L} \le L_{\max}$. Then, we have*

$$\begin{aligned}\|\hat{\lambda} - \lambda_0\|^2 \le \inf_{\beta\in\mathbb{R}^m_+} \big\|\lambda_\beta - \lambda_0\big\|^2 + 6(L_{\max} + 2(L_0-1)) \max_{1\le j\le m} \hat{w}_j^2 \\ + K_1 \frac{\|\lambda_0\|_\infty \big(x + L_{\max}(1+\log m)\big)}{n} \\ + K_2 \frac{m\big(x + L_{\max}(1+\log m)\big)^2}{n^2},\end{aligned} \tag{2.11}$$

with a probability larger than $1 - L_{\max} e^{-x}$, with $\|\lambda_0\|_\infty = \sup_{t\in[0,1]} \lambda_0(t)$, $K_1 = 1670.89$, and $K_2 = 6683.53$.

The proof of Theorem 2.3.3 is provided in Section 2.6. This results proves that our procedure has a fast rate of convergence of order

$$\frac{(L_{\max} \vee L_0) m \log m}{n},$$

which scales in m/n.

Corollary 2.3.4. *Given Assumption 2.2.1, and under the same assumptions as the ones from Theorem 2.3.3, we have*

$$\begin{aligned}\|\hat{\lambda} - \lambda_0\|^2 \le \frac{2(L_0-1)\Delta^2_{\beta,\max}}{m} + 6(L_{\max} + 2(L_0-1)) \max_{1\le j\le m} \hat{w}_j^2 \\ + K_1 \frac{\|\lambda_0\|_\infty \big(x + L_{\max}(1+\log m)\big)}{n} \\ + K_2 \frac{m\big(x + L_{\max}(1+\log m)\big)^2}{n^2},\end{aligned} \tag{2.12}$$

with a probability larger than $1 - L_{\max} e^{-x}$, with the same notations as in Theorem 2.3.3.

The proof of Corollary 2.3.4 is presented in Section 2.6. A consequence of Corollary 2.3.4 is that an optimal tradeoff between approximation and complexity is given by the choice $m \approx n^{1/2}$. Note that we are able to use the same procedure in Theorems 2.3.1 and 2.3.3, namely for the slow and fast rate, while it is not the case in the signal + white noise considered in Harchaoui and Lévy-Leduc (2010) for instance.

2.4 Change-point detection

In this section we prove that the proposed total-variation with data-driven weights procedure is consistent for the estimation of the change-point positions. Note that, however, the context considered here is quite different from the more standard signal + white noise setting: here we aim at detecting change-points in the intensity function, hence this problem is prone to an unavoidable non-parametric bias of approximation by a piecewise constant function. This means that we will not be able to recover the exact position of two change-points if they lie on the same interval $I_{j,m}$. Therefore, we assume

Assumption 2.4.1. *Grant Assumption 2.2.1 and assume that there is a positive constant* $c \geq 8$ *such that*

$$\min_{1 \leq \ell \leq L_0} |\tau_{0,\ell} - \tau_{0,\ell-1}| > \frac{c}{m}. \tag{2.13}$$

This assumption entails that the change-points of λ_0 are sufficiently far apart, and that, in particular, there cannot be more than one change-point in the "high-resolution" intervals $I_{j,m}$. Under Assumption 2.4.1, the procedure will be able to recover the (unique) intervals $I_{j_\ell,m}$, for $\ell = 0,\dots,L_0$, where the change-point belongs. Hence, we define the *approximate change-points sequence* $[j_\ell]_{0 \leq \ell \leq L_0}$ as follows.

Definition 2.4.2. The *approximate change-points sequence* $[j_\ell]_{0 \leq \ell \leq L_0}$ relative to the level of resolution m is defined as the right-hand side boundary of the unique interval $I_{j_\ell,m}$ that contains the change-point $\tau_{0,\ell}$, namely

$$\tau_{0,\ell} \in \left(\frac{j_\ell - 1}{m}, \frac{j_\ell}{m} \right] \tag{2.14}$$

for $\ell = 1,\dots,L_0 - 1$, where we put $j_0 = 0$ and $j_{L_0} = m$ by convention.

Given the support $\hat{S} = \{\hat{j}_1,\dots,\hat{j}_{\hat{L}}\}$ with $\hat{j}_1 < \cdots < \hat{j}_{\hat{L}}$ of the discrete gradient of $\hat{\beta}$ defined in (2.7), and introducing $\hat{j}_0 = 0$ and $\hat{j}_{\hat{L}+1} = m$, we define simply

$$\hat{\tau}_\ell = \frac{\hat{j}_\ell}{m} \tag{2.15}$$

for $\ell = 0,\dots,\hat{L}+1$. In order to be able to prove a consistency results for change-points detection, we need a set of assumptions that quantifies the asymptotic interplay between several quantities:

— $\Delta_{j,\min} = \min_{1 \leq \ell \leq L_0 - 1} |j_{\ell+1} - j_\ell|$, which is the minimum distance between two consecutive terms in the change-points of λ_0.

— $\Delta_{\beta,\min} = \min_{1\le q\le m-1} |\beta_{0,q+1,m} - \beta_{0,q,m}|$, which is the smallest jump size of the projection $\lambda_{0,m}$ of λ_0 onto Λ_m.
— $(\varepsilon_n)_{n\ge 1}$, a non-increasing and positive sequence that goes to zero as $n \to \infty$, and such that $m\varepsilon_n \ge 6$ for any $n \ge 1$.

Assumption 2.4.3. *We assume that $\Delta_{j,\min}$, $\Delta_{\beta,\min}$ and $(\varepsilon_n)_{n\ge 1}$ satisfy*

$$\frac{\sqrt{nm}\varepsilon_n \Delta_{\beta,\min}}{\sqrt{\log m}} \to \infty \tag{2.16}$$

$$\frac{\sqrt{n}\Delta_{j,\min}\Delta_{\beta,\min}}{\sqrt{m\log m}} \to \infty \tag{2.17}$$

as $n \to \infty$.

This assumption controls the rate (ε_n) of convergence of $\hat{\tau}_\ell$ towards $\tau_{0,\ell}$. The logarithmic factor is due to concentration inequalities for the control of the noise (the martingale M obtained by compensation of N). The next Theorem proves the consistency of our procedure for the detection of change-points, under the assumption that the estimated number of change-points is the correct one.

Theorem 2.4.4. *Under Assumptions 2.4.1 and 2.4.3, and if $\hat{L} = L_0 - 1$, then the change-points estimators $\{\hat{\tau}_1, \dots, \hat{\tau}_{\hat{L}}\}$ given by (2.15) satisfy*

$$\mathbb{P}\Big[\max_{1\le \ell \le L_0 - 1} |\tau_{0,\ell} - \hat{\tau}_\ell| \le \varepsilon_n\Big] \to 1 \tag{2.18}$$

as $n \to \infty$.

The proof of Theorem 2.4.4 is quite involved and is presented in Section 2.7 and Section 2.B. It builds upon some techniques developed in Harchaoui and Lévy-Leduc (2010), based on a careful inspection of the Karush-Kuhn-Tucker (KKT) optimality conditions, see for instance Boyd and Vandenberghe (2004), for the solutions to the convex problem (2.6). The proof depends also heavily on a data-driven Bernstein's inequality for the control of the martingale errors, see Proposition 2.6.1 from Section 2.6.

Let us give examples of scaling for the quantities $\Delta_{j,\min}$, $\Delta_{\beta,\min}$ and $(\varepsilon_n)_{n\ge 1}$ that meet Assumption 2.4.3. Assume for simplicity that

$$\varepsilon_n = n^{-\alpha} \quad \text{and} \quad \Delta_{\beta,\min} = n^{-\gamma}$$

for some constants $\alpha, \gamma > 0$.

— If $m = n^{1/3}$ then Theorem 2.4.4 holds with any $\alpha, \gamma > 0$ satisfying $0 < \gamma < 1/3$ and $0 < \alpha + \gamma < 2/3$, and if $\Delta_{j,\min} \ge 6$.
— If $m = n^{1/2}$ then Theorem 2.4.4 holds with any $0 < \gamma < 1/4$ and $0 < \alpha + \gamma < 3/4$ and if $\Delta_{j,\min} \ge 6$.

In order to prove change-point consistency without the assumption that the estimated number of change-points is the correct one, we need to relax a little bit the statement of the result given in Theorem 2.4.4. Namely, we evaluate a non-symmetrized Hausdorf distance $\mathscr{E}(\hat{\mathcal{T}} \| \mathcal{T}_0)$ between the set of estimated change-points

$$\hat{\mathcal{T}} = \{\hat{\tau}_1, \dots, \hat{\tau}_{\hat{L}}\}$$

and the set of true change-points

$$\mathcal{T}_0 = \{\tau_{0,1}, \ldots, \tau_{0,L_0-1}\},$$

where for two sets A and B, the quantity $\mathscr{E}(A\|B)$ is given by

$$\mathscr{E}(A\|B) = \sup_{b\in B} \inf_{a\in A} |a-b|.$$

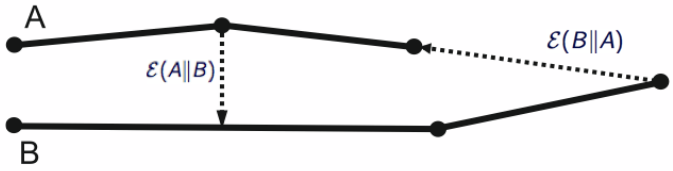


Fig. 2.1 – Hausdorff distance between A and B.

Note that $\mathscr{E}(A\|B) \vee \mathscr{E}(B\|A)$ is the Hausdorff distance between A and B, see Figure 2.1. When $\hat{L} = L_0 - 1$, Theorem 2.4.4 implies that

$$\mathbb{P}\left[\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0) \le \varepsilon_n, \mathscr{E}(\mathcal{T}_0\|\hat{\mathcal{T}}) \le \varepsilon_n\right] \to 1 \tag{2.19}$$

as $n \to \infty$. When $\hat{L} > L_0 - 1$, we prove in Theorem 2.4.5 below that $\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0) \le \varepsilon_n$ with a probability going to 1 as $n \to \infty$. This means that change-point consistency holds for our procedure whenever the estimated number of change-points is not less than the true one.

Theorem 2.4.5. *Under Assumption 2.4.1 and Assumption 2.4.3, and if $\hat{L} \ge L_0 - 1$, we have*

$$\mathbb{P}\left[\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0) \le \varepsilon_n\right] \to 1 \tag{2.20}$$

as $n \to \infty$.

Theorem 2.4.5 ensures that even when the number of change-points is over-estimated, each true change-point is close to the estimated one. The proof of Theorem 2.4.5 is given in Section 2.8. It is based, as for the proof of Theorem 2.4.4, on a repeated utilization of the KKT optimality conditions of problem (2.6).

Note that a difference with Harchaoui and Lévy-Leduc (2010) is that we are able to use the same regularization parameters $\hat{w}_j$ given by (2.10) in Theorems 2.4.4 and 2.4.5. Besides, we don't need an upper bound on the estimated number of change-points in Theorem 2.4.5, while it is necessary in Qian and Su. (2013).

2.5 Numerical experiments

In this section we propose a fast algorithm for solving the optimization problem (2.6) and apply it on simulated and real datasets from genomics.

2.5.1 Algorithm

A concept of importance for convex optimization in machine learning is the proximal operator see Bach et al. (2012) and Bauschke and Combettes (2011). The proximal operator prox_f of a proper, lower semicontinuous, convex function $f : \mathbb{R}^m \to (-\infty, \infty]$, is defined as

$$\text{prox}_f(v) = \text{argmin}_{x \in \mathbb{R}^m} \left\{ \frac{1}{2}\|v - x\|_2^2 + f(x) \right\}, \text{ for all } v \in \mathbb{R}^m.$$

In this section, we provide a fast algorithm to solve the optimization problem (2.6), that computes the proximal operator of the weighted total-variation.

We observe n i.i.d observations of N over the interval $[0,1]$. Recall that $\bar{N}_n = n^{-1}\sum_{i=1}^n N_i$, and $\bar{N}_n(I) = \int_I d\bar{N}_n(t)$ for any $I \subset [0,1]$. We also recall that $\hat{\lambda}(t) = \sum_{j=1}^m \hat{\beta}_j \lambda_{j,m}(t)$, where $\hat{\beta} = [\hat{\beta}_1, \ldots, \hat{\beta}_m]$ is given by (2.6). Hence, we have

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}_+^m} \left\{ \frac{1}{2}\|\mathbf{N} - \beta\|_2^2 + \|\beta\|_{\text{TV},\hat{w}} \right\}, \tag{2.21}$$

where $\mathbf{N} = [\mathbf{N}_j]_{1 \le j \le m} \in \mathbb{R}_+^m$ is given by

$$\mathbf{N} = \begin{bmatrix} \sqrt{m}\bar{N}_n(I_{1,m}) \\ \vdots \\ \sqrt{m}\bar{N}_n(I_{m,m}) \end{bmatrix}.$$

Therefore, we see that (2.21) is equivalent to

$$\hat{\beta} = \text{prox}_{\|\cdot\|_{\text{TV},\hat{w}}}(\mathbf{N}).$$

Next, we develop an algorithm that computes $\text{prox}_{\|\cdot\|_{\text{TV},\hat{w}}}$, which is an extension of Condat (2013) to weighted total-variation. Towards this end, we introduce the following $(m-1) \times m$ bidiagonal matrix

$$D_{\hat{w}} = \begin{bmatrix} -\hat{w}_2 & \hat{w}_2 & 0 & \cdots & 0 \\ 0 & -\hat{w}_3 & \hat{w}_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\hat{w}_m & \hat{w}_m \end{bmatrix}.$$

Then, one can express the primal problem (2.21) as follows:

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}_+^m} \left\{ \frac{1}{2}\|\mathbf{N} - \beta\|_2^2 + \|D_{\hat{w}}\beta\|_1 \right\}. \tag{2.22}$$

Essentially, problem (2.22) is difficult to analyse directly because the nondifferentiable ℓ_1 norm is composed with a linear transformation of β. When solving (2.22) we may consider its Fenchel dual form Bauschke and Combettes (2011). First, we rewrite the primal problem as

$$\begin{aligned} &\text{minimize}_{\beta \in \mathbb{R}^m, z \in \mathbb{R}^{m-1}} \frac{1}{2}\|\mathbf{N} - \beta\|_2^2 + \|z\|_1 \\ &\text{subject to } D_{\hat{w}}\beta = z, \end{aligned}$$

whose Lagrangian is

$$\mathscr{L}(\beta, z, u) = \frac{1}{2}\|\mathbf{N} - \beta\|_2^2 + \|z\|_1 + u^\top (D_{\hat{w}}\beta - z),$$

and to derive a dual problem, we minimize this over β, z. A straightforward computation gives

$$\min_{\beta}\left\{\frac{1}{2}\|\mathbf{N} - \beta\|_2^2 + u^\top D_{\hat{w}}\beta\right\} = -\frac{1}{2}\|\mathbf{N} - D_{\hat{w}}^\top u\|_2^2,$$

while

$$\min_{z}\left\{\|z\|_1 - u^\top z\right\} = \begin{cases} 0, & \text{if } \|u\|_\infty \le 1, \\ -\infty, & \text{otherwise.} \end{cases}$$

Introducing $u_0 = u_m = 0$, we proved that a dual problem of (2.22) is given by

$$\begin{aligned} &\text{minimize}_{u\in\mathbb{R}^{m+1}} \frac{1}{2}\sum_{k=1}^{m}\left(\mathbf{N}_k - \hat{w}_{k+1}u_k + \hat{w}_k u_{k-1}\right)^2, \\ &\text{subject to } |u_j| \le 1, \text{ for } k = 1, \dots, m, \text{ and } u_0 = u_m = 0. \end{aligned}$$

If we have a feasible dual variable $\hat{u}$, we can compute the primal solution $\hat{\beta}$ using

$$\hat{\beta}_k = \mathbf{N}_k - \hat{w}_{k+1}\hat{u}_k + \hat{w}_k\hat{u}_{k-1}, \text{ for } k = 1, \dots, m. \tag{2.23}$$

For this problem, strong duality holds, see Boyd and Vandenberghe (2004), meaning that the duality gap is zero. The KKT optimality conditions characterize the unique solutions $\hat{\beta}$ and $\hat{\theta}_k := \hat{w}_{k+1}\hat{u}_k$. They yield, in addition to (2.23):

$$\hat{\theta}_0 = \hat{\theta}_m = 0, \text{ and } \forall\, k = 1, \dots, m-1, \begin{cases} \hat{\theta}_k \in [-\hat{w}_{k+1}, \hat{w}_{k+1}], & \text{if } \hat{\beta}_k = \hat{\beta}_{k+1}, \\ \hat{\theta}_k = -\hat{w}_{k+1}, & \text{if } \hat{\beta}_k < \hat{\beta}_{k+1}, \\ \hat{\theta}_k = \hat{w}_{k+1}, & \text{if } \hat{\beta}_k > \hat{\beta}_{k+1}. \end{cases} \tag{2.24}$$

Therefore, the proposed algorithm consists in running forwardly through the samples $[\mathbf{N}_k]_{1\le k\le m}$. Using (2.24), at location k, $\hat{\beta}_k$ stays constant where $|\hat{\theta}_k| < \hat{w}_{k+1}$. If this is not possible, it goes back to the last location where a jump can be introduced in $\hat{\beta}$, validates the current segment until this location, starts a new segment, and continues. This algorithm is described precisely in Algorithm 3.

2.5.2 Simulated data

We conduct simulations on 2 examples of intensities. We simulate counting processes with inhomogeneous piecewise intensities λ_0, with 5, and 15 change points, see Figure 2.2, with an increasing sample size n. In order to assess the performance of the total-variation procedure $\hat{\lambda}$, we use a Monte-Carlo averaged mean integrated squared error (MISE) as a performance measure, given by

$$\text{MISE}(\hat{\lambda}, \lambda_0) = \mathbb{E}\int_0^1 (\hat{\lambda}(t) - \lambda_0(t))^2 dt.$$

We run 100 Monte-Carlo experiments, for an increasing sample size between $n = 500$ and $n = 30000$, for each 2 examples. In Figure 2.3, we plot the MISEs of the weighted and the unweighted total-variation (namely $\hat{w} \equiv 1$), for the 3 examples, as a function of

Algorithm 3: $\hat{\beta} = \mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}(\mathbf{N})$

Input: $\mathbf{N} = \left(\mathbf{N}_1, \ldots, \mathbf{N}_m\right)^\top \in \mathbb{R}^m$; $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m) \in \mathbb{R}^m_+$.

Output: $\left(\hat{\beta}_1, \ldots, \hat{\beta}_m\right)^\top$.

1. **Set** $k = k_0 = k_- = k_+ \leftarrow 1$;
   $\beta_{\min} \leftarrow \mathbf{N}_1 - \hat{w}_2$; $\beta_{\max} \leftarrow \mathbf{N}_1 + \hat{w}_2$;
   $\theta_{\min} \leftarrow \hat{w}_2$; $\theta_{\max} \leftarrow -\hat{w}_2$;
2. **if** $k = m$ **then**
   $\hat{\beta}_m \leftarrow \beta_{\min} + \theta_{\min}$;
3. **if** $\mathbf{N}_{k+1} + \theta_{\min} < \beta_{\min} - \hat{w}_{k+2}$ **then** /* negative jump */
   $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_-} \leftarrow \beta_{\min}$;
   $k = k_0 = k_- = k_+ \leftarrow k_- + 1$;
   $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{w}_{k+1} + \hat{w}_k$; $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{w}_{k+1} + \hat{w}_k$;
   $\theta_{\min} \leftarrow \hat{w}_{k+1}$; $\theta_{\max} \leftarrow -\hat{w}_{k+1}$;
4. **else if** $\mathbf{N}_{k+1} + \theta_{\max} > \beta_{\max} + \hat{w}_{k+2}$ **then** /* positive jump */
   $\hat{\beta}_{k_0} = \ldots = \hat{\beta}_{k_+} \leftarrow \beta_{\max}$;
   $k = k_0 = k_- = k_+ \leftarrow k_+ + 1$;
   $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{w}_{k+1} - \hat{w}_k$; $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{w}_{k+1} - \hat{w}_k$;
   $\theta_{\min} \leftarrow \hat{w}_{k+1}$; $\theta_{\max} \leftarrow -\hat{w}_{k+1}$;
5. **else** /* no jump */
   set $k \leftarrow k + 1$;
   $\theta_{\min} \leftarrow \mathbf{N}_k + \hat{w}_{k+1} - \beta_{\min}$;
   $\theta_{\max} \leftarrow \mathbf{N}_k - \hat{w}_{k+1} - \beta_{\max}$;
   if $\theta_{\min} \geq \hat{w}_{k+1}$ **then**
   $\beta_{\min} \leftarrow \beta_{\min} + \frac{\theta_{\min} - \hat{w}_{k+1}}{k - k_0 + 1}$;
   $\theta_{\min} \leftarrow \hat{w}_{k+1}$;
   $k_- \leftarrow k$;
   if $\theta_{\max} \leq -\hat{w}_{k+1}$ **then**
   $\beta_{\max} \leftarrow \beta_{\max} + \frac{\theta_{\max} + \hat{w}_{k+1}}{k - k_0 + 1}$;
   $\theta_{\max} \leftarrow -\hat{w}_{k+1}$;
   $k_+ \leftarrow k$;
6. **if** $k < m$ **then**
   go to **3.**;
7. **if** $\theta_{\min} < 0$ **then**
   $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_-} \leftarrow \beta_{\min}$;
   $k = k_0 = k_- \leftarrow k_- + 1$;
   $\beta_{\min} \leftarrow \mathbf{N}_k - \hat{w}_{k+1} + \hat{w}_k$;
   $\theta_{\min} \leftarrow \hat{w}_{k+1}$; $\theta_{\max} \leftarrow \mathbf{N}_k + \hat{w}_k - v_{\max}$;
   go to **2.**;
8. **else if** $\theta_{\max} > 0$ **then**
   $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_{k_+} \leftarrow \beta_{\max}$;
   $k = k_0 = k_+ \leftarrow k_+ + 1$;
   $\beta_{\max} \leftarrow \mathbf{N}_k + \hat{w}_{k+1} - \hat{w}_k$;
   $\theta_{\max} \leftarrow -\hat{w}_{k+1}$; $\theta_{\min} \leftarrow \mathbf{N}_k - \hat{w}_k - \theta_{\min}$;
   go to **2.**;
9. **else**
   $\hat{\beta}_{k_0} = \cdots = \hat{\beta}_m \leftarrow \beta_{\min} + \frac{\theta_{\min}}{k - k_0 + 1}$;

the sample size. We observe in Figure 2.3 that the estimation error is always decaying with the sample size, and that both procedures behave similarly. Differences can be observed below, using a genomics datasets. On each simulated dataset, we perform a 10-fold cross-validation to select the best constant to use in front of the weights $\hat{w}_j$ (both for the weighted and unweighted total-variation). Cross-validation in this context is achieved by choosing uniformly at random a label between 1 and 10 for each point, and by using points with label k in the k-th testing fold and removing these points for the k-th training fold. The estimated intensity is accordingly corrected, by this amount (as removing uniformly a fraction of points from a counting process biases downwards the intensity by the same fraction).

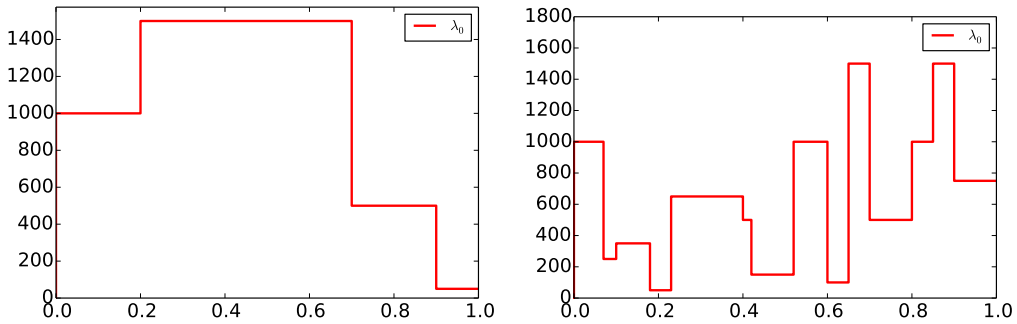


Fig. 2.2 – Intensities used for Example 1 (left), and Example 2 (right) respectively with 5, and 15 change-points.

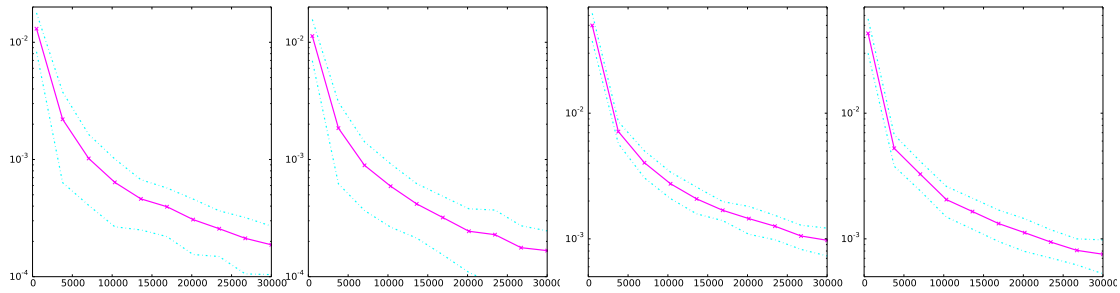


Fig. 2.3 – Average MISEs (bold lines) over 100 Monte-Carlo experiments and standard deviations of the MISEs (dashed lines). First: weighted TV for Example 1; Second: non-weighted TV for Example 1; Third: weighted TV for Example 2; Fourth: non-weighted TV for Example 2.

2.5.3 Real data

Our method is illustrated on NCI-60 tumor and normal cell lines, HCC1954 and BL1954. This dataset was produced and investigated by Chiang et al. (2009) using the Illumina platform, where the reads are 36bp long. After cleaning of this data, there are 7.72 million reads for the tumor (HCC1954) and 6.65 million reads for the normal (BL1954) samples respectively. A description of the sampling process for such data is described in Introduction. We show in Figures 2.4 and 2.5 both tumor and

cell lines data. This data consists of a list of reads number, see Figure 2.5, where we plot a zoomed sequence of reads. For visualization purposes, we give in Figure 2.5 the binned counts of reads over 10000 intervals equispaced on the range of reads.

Fig. 2.4 – A zoom into the sequence of reads for normal (left) and tumor (right) data.

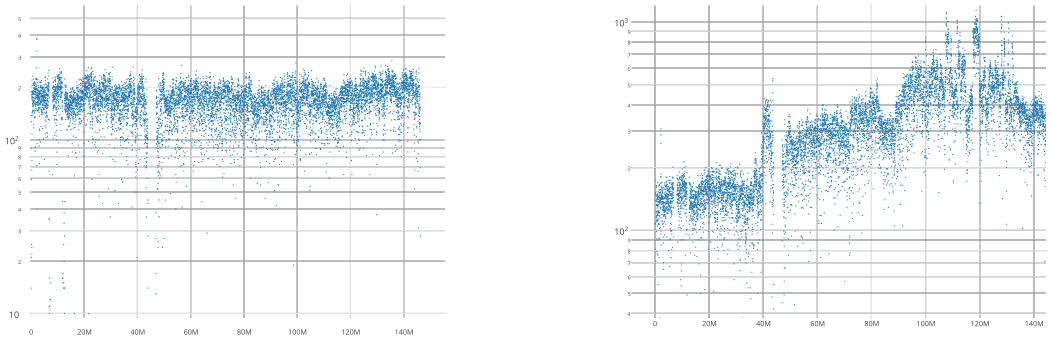


Fig. 2.5 – Binned counts of reads (log-scale) of the normal (left) and tumor (right) data.

In Figure 2.6 we plot the best solution of the weighted and unweighted ($\hat{w}_j = 1$) total-variation estimators on the normal and tumor reads data. For easier visualization we plot a zoom of the reads sequence. We perform a 10-fold cross-validation to select the best constant to use in front of the weights $\hat{w}_j$ (both for the weighted and unweighted total-variation), as explained above. We observe in this figure that the weighted total-variation gives sharper results: the piecewise constant intensity is smoother, and the obtained change-points locations seem, at least visually, better. An important fact is that the runtime of Algorithm 3 is extremely fast: a solution is obtained in less than one millisecond, on a modern laptop (implementation is done using python with a C extension). This is due to the fact that Algorithm 3 is typically linear in the signal size.

2.6 Proof of Theorems 2.3.1 and 2.3.3

Introduce $\mu = [\mu_j]_{1\leq j\leq m} \in \mathbb{R}^m$ given by $\mu_1 = \beta_1$ and $\mu_j = \beta_j - \beta_{j-1}$ for $j = 2,\dots,m$. Then, we have $\beta = \mathbf{T}\mu$, where $\mathbf{T}$ is the $m \times m$ lower triangular matrix with entries $(\mathbf{T})_{j,k} = 0$ if $j < k$ and $(\mathbf{T})_{j,k} = 1$ otherwise. Note that $\hat{\beta} = \mathbf{T}\hat{\mu}$, where

$$\hat{\mu} = \operatorname{argmin}_{\mu\in\mathbb{R}^m} \Big\{ \frac{1}{2}\|\mathbf{N} - \mathbf{T}\mu\|_2^2 + \sum_{j=2}^{m} \hat{w}_j |\mu_j| \Big\}. \tag{2.25}$$

2.6.1 Proof of Theorem 2.3.1

This proof follows a standard argument for proving slow oracle inequalities, see for instance Bickel et al. (2009). Due to the Doob-Meyer decomposition theorem, we

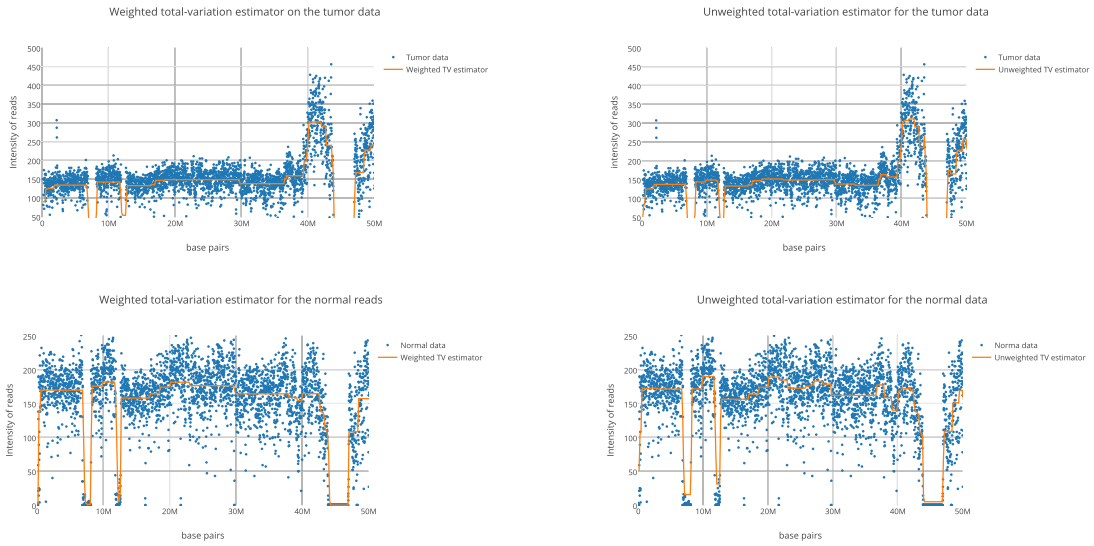


Fig. 2.6 – A zoom between reads number 0 and 50M of the weighted and unweightedtotal-variation estimators applied to the tumor and normal data.

have

$$R_n(\lambda) = \|\lambda - \lambda_0\|^2 - \|\lambda_0\|^2 - \int_0^1 \lambda(t) d\bar{M}_n(t),$$

which leads to

$$\hat{\lambda} = \lambda_{\hat{\beta}} = \operatorname{argmin}_{\beta \in \mathbb{R}_+^m} \left(\|\lambda_\beta - \lambda_0\|^2 - 2 \int_0^1 \lambda_\beta(t) d\bar{M}_n(t) + \|\beta\|_{\mathrm{TV},\hat{w}} \right). \tag{2.26}$$

Then, using (2.6), it implies that

$$\|\hat{\lambda} - \lambda_0\|^2 \le \inf_\beta \|\lambda_\beta - \lambda_0\|^2 + \frac{2}{n} \nu_n(\hat{\lambda} - \lambda_\beta) + \|\beta\|_{\mathrm{TV},\hat{w}} - \|\hat{\beta}\|_{\mathrm{TV},\hat{w}}, \tag{2.27}$$

where $\nu_n(\lambda) = \sum_{i=1}^n \int_0^1 \lambda(t) dM_i(t)$ is a centered empirical process. Note that

$$\begin{aligned} \frac{1}{n} \nu_n(\hat{\lambda} - \lambda_\beta) &= \sum_{j=1}^m (\hat{\beta}_{j,m} - \beta_{j,m}) \int_0^1 \lambda_{j,m}(t) d\bar{M}_n(t) \\ &= \sum_{j=1}^m ((\mathbf{T}\hat{\mu})_{j,m} - (\mathbf{T}\mu)_{j,m}) \int_0^1 \lambda_{j,m}(t) d\bar{M}_n(t) \\ &= \sum_{j=1}^m (\hat{\mu}_{j,m} - \mu_{j,m}) \sum_{q=j}^m \int_0^1 \lambda_{q,m}(t) d\bar{M}_n(t). \end{aligned} \tag{2.28}$$

Define the event Ω_n by

$$\Omega_n = \bigcap_{j=1}^m \left\{ \left| \sum_{q=j}^m \int_0^1 \lambda_{q,m}(t) d\bar{M}_n(t) \right| \le \frac{\hat{w}_j}{2} \right\}.$$

The probabilistic control of Ω_n is given in Proposition 2.6.1 from Section 2.6 below. It relies on a slight modification of an empirical Bernstein inequality from Gaïffas and Guilloux (2012), see also Reynaud-Bouret (2003). On Ω_n, we have using (2.28)

$$\frac{2}{n} \nu_n(\hat{\lambda} - \lambda_\beta) \le \sum_{j=1}^m \hat{w}_j |\hat{\mu}_{j,m} - \mu_{j,m}|,$$

Using (2.27), we obtain

$$\begin{aligned}\|\hat{\lambda}-\lambda_0\|^2 &\le \|\lambda_\beta-\lambda_0\|^2+\sum_{j=1}^m \hat{w}_j|\hat{\mu}_{j,m}-\mu_{j,m}|+\sum_{j=1}^m \hat{w}_j(|\mu_{j,m}|-|\hat{\mu}_{j,m}|)\\ &\le \|\lambda_\beta-\lambda_0\|^2+2\sum_{j=1}^m \hat{w}_j|\mu_{j,m}|\\ &= \|\lambda_\beta-\lambda_0\|^2+2\|\beta\|_{\mathrm{TV},\hat{w}}.\end{aligned}$$

Then, on Ω_n, (2.8) in Theorem 2.3.1 holds true . It remains now to control $\mathbb{P}(\Omega_n^\complement)$. We have, recalling $\lambda_{j,m}(t)=\sqrt{m}\mathbb{1}_{(\frac{j-1}{m},\frac{j}{m}]}(t)$, that

$$\mathbb{P}[\Omega_n^\complement]\le\sum_{j=1}^m \mathbb{P}\Big[\Big|\sqrt{m}\int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(t)d\bar{M}_n(t)\Big|>\frac{\hat{w}_j}{2}\Big],$$

so we need to control the tails of

$$U_j=\int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(t)d\bar{M}_n(t),$$

which is the goal of the next proposition.

Proposition 2.6.1. *For any numerical constants $c_h>1$, $\varepsilon>0$ and $c_0>0$ such that $ec_0>2(4/3+\varepsilon)c_h$, the following holds for any $z>0$:*

$$\mathbb{P}\Big[|U_j|\ge c_{1,\varepsilon}\sqrt{\frac{z+\hat{h}_{n,z,j}}{n}\hat{V}_j}+c_{3,\varepsilon}\frac{z+1+\hat{h}_{n,z,j}}{n}\Big]\le ce^{-z}$$

where

$$\hat{h}_{n,z,j}=c_h\log\log\Big(\frac{2en\hat{V}_j+2e(\frac{4}{3}+\varepsilon)z}{ec_0(z+1)-2(\frac{4}{3}+\varepsilon)c_h}\vee e\Big),$$

$c_{1,\varepsilon}=2\sqrt{1+\varepsilon}, c_{3,\varepsilon}=\sqrt{2\max\big(c_0,2(1+\varepsilon)(\frac{4}{3}+\varepsilon)\big)}+\frac{1}{3}$, *and* $c=6+4\big(\log(1+\varepsilon)\big)^{-c_h}\sum_{q\ge1}q^{-c_h}$.

The proof of Proposition 2.6.1 is given in Section 2.A.1. Choosing $z=x+\log m$, it yields that

$$\begin{aligned}&\sum_{j=1}^m \mathbb{P}\Big[|U_j|\ge c_{1,\varepsilon}\sqrt{\frac{x+\log m+\hat{h}_{n,x,j}}{n}\hat{V}_j}+c_{3,\varepsilon}\frac{x+\log m+\hat{h}_{n,x,j}+1}{n}\Big]\\ &\le\big(6+4\big(\log(1+\varepsilon)\big)^{-c_h}\sum_{q\ge1}q^{-c_h}\big)e^{-x},\end{aligned}$$

where

$$\hat{h}_{n,x,j}=c_h\log\log\Big(\frac{2en\hat{V}_j+2e(\frac{4}{3}+\varepsilon)(x+\log m)}{ec_0(x+\log m+1)-2(\frac{4}{3}+\varepsilon)c_h}\vee e\Big).$$

Then, the choice of data-driven weights is given by

$$\hat{w}_j=c_1\sqrt{\frac{m(x+\log m+\hat{h}_{n,x,j})\hat{V}_j}{n}}+c_2\frac{\sqrt{m}(x+1+\log m+\hat{h}_{n,x,j})}{n},$$

where $c_1=2c_{1,\varepsilon}$ and $c_2=2c_{3,\varepsilon}$ gives $\mathbb{P}(\Omega_n^\complement)\le ce^{-x}$. Finally, to get the numerical constants in Theorem 2.3.1, we set $\varepsilon=1, c_h=2$, and $c_0=28/3e$ in Proposition 2.6.1. □

2.6.2 Proof of Corollary 2.3.2

We denote by $\lambda_{0,m}$ the projection of λ_0 onto Λ_m, that is $\lambda_{0,m} = \operatorname{argmin}_{\lambda_\beta \in \Lambda_m} \|\lambda_\beta - \lambda_0\|^2$. Using Pythagoras' theorem, we have

$$\|\hat{\lambda} - \lambda_0\|^2 \leq \|\lambda_{0,m} - \lambda_0\|^2 + \|\hat{\lambda} - \lambda_{0,m}\|^2.$$

By the proof of Theorem 2.3.1, we obtain

$$\begin{aligned} \|\hat{\lambda} - \lambda_{0,m}\|^2 &\leq 2\|\beta_{0,m}\|_{\mathrm{TV},\hat{w}} \\ &\leq 2\|\beta_{0,m}\|_{\mathrm{TV}} \max_{1\leq j\leq m} \hat{w}_j. \end{aligned}$$

Now, the following approximation lemma comes in handy for the control of the bias term.

Lemma 2.6.2. *Given Assumption 2.2.1, we have*

$$\|\lambda_{0,m} - \lambda_0\|^2 \leq \frac{2(L_0 - 1)\Delta_{\beta,\max}^2}{m},$$

where $\Delta_{\beta,\max} = \max_{1\leq \ell,\ell' \leq L_0} |\beta_{0,\ell} - \beta_{0,\ell'}|$.

The proof of Lemma 2.6.2 is given in Section 2.A.2. □

2.6.3 Proof of Theorem 2.3.3

Using Pythagoras' identity, we obtain the following decomposition

$$\|\lambda_{\hat{\beta}} - \lambda_0\|^2 = \|\lambda_\beta - \lambda_0\|^2 + \|\lambda_{\hat{\beta}} - \lambda_\beta\|^2.$$

In view of the fact that $\{\lambda_{j,m} : j = 1,\dots,m\}$ is an orthonormal basis of Λ_m, we have

$$\|\lambda_{\hat{\beta}} - \lambda_\beta\|^2 = \|\hat{\beta} - \beta\|_2^2,$$

and by the definition of $\hat{\beta}$, we get

$$\|\hat{\beta} - \mathbf{N}\|_2^2 + \sum_{j=2}^m \hat{w}_j |\hat{\beta}_{j,m} - \hat{\beta}_{j-1,m}| \leq \|\beta - \mathbf{N}\|_2^2 + \sum_{j=2}^m \hat{w}_j |\beta_{j,m} - \beta_{j-1,m}|.$$

Then

$$\|\hat{\beta} - \beta\|_2^2 \leq \sum_{j=2}^m \hat{w}_j \Big(|\beta_{j,m} - \beta_{j-1,m}| - |\hat{\beta}_{j,m} - \hat{\beta}_{j-1,m}| \Big) + 2\int_0^1 \sum_{j=2}^m (\hat{\beta}_{j,m} - \beta_{j,m}) \lambda_{j,m}(t) d\bar{M}_n(t).$$

Assume that $\hat{\beta}$ belongs to a set of dimension at most $L_{\max}$. Let $S = \{j : \beta_{j,m} \neq \beta_{j-1,m} \text{ for } j = 2,\dots,m\}$, be the support of the discrete gradient of β. Using the Cauchy–Schwarz in-

equality, we have

$$\begin{aligned}
&\sum_{j=2}^{m} \hat{w}_j \Big(|\beta_{j,m} - \beta_{j-1,m}| - |\hat{\beta}_{j,m} - \hat{\beta}_{j-1,m}| \Big) \\
&\quad \le \sum_{j \in \hat{S} \cup S} \hat{w}_j \Big(|\beta_{j,m} - \hat{\beta}_{j,m}| + |\beta_{j-1,m} - \hat{\beta}_{j-1,m}| \Big) \\
&\quad \le \sum_{j \in \hat{S} \cup S} \hat{w}_j \Big(|\beta_{j,m} - \hat{\beta}_{j,m}| \Big) + \sum_{j \in \hat{S} \cup S} \hat{w}_j \Big(|\beta_{j-1,m} - \hat{\beta}_{j-1,m}| \Big) \\
&\quad \le \sum_{j \in \hat{S} \cup S \cup (\hat{S} \cup S+1)} \hat{w}_j \Big(|\hat{\beta}_{j,m} - \beta_{j,m}| \Big) \\
&\quad \le \sqrt{|\hat{S} \bigcup S \bigcup (\hat{S}+1) \bigcup (S+1)|} \\
&\qquad\qquad \times \Big\| \Big[\hat{\beta}_{j,m} - \beta_{j,m} \Big]_{j \in \hat{S} \cup S \cup (\hat{S}+1) \cup (S+1)} \Big\|_2 \times \max_{j \in \hat{S} \cup S \cup (\hat{S}+1) \cup (S+1)} \hat{w}_j \\
&\quad \le \sqrt{2} \sqrt{L_{\max} + 2(L_0 - 1)} \, \big\| \hat{\beta} - \beta \big\|_2 \max_{j=1,\dots,m} \hat{w}_j .
\end{aligned}$$

Hence

$$\begin{aligned}
\big\| \hat{\beta} - \beta \big\|_2^2 \le \sqrt{2} \sqrt{L_{\max} + 2(L_0 - 1)} \, \big\| \hat{\beta} - \beta \big\|_2 \max_{j=1,\dots,m} \hat{w}_j \\
+ 2 \big\| \hat{\beta} - \beta \big\|_2 \int_0^1 \sum_{j=2}^{m} \frac{(\hat{\beta}_{j,m} - \beta_{j,m}) \lambda_{j,m}(t)}{\big\| \hat{\beta} - \beta \big\|_2} d\bar{M}_n(t).
\end{aligned}$$

Now, define the functional G for all $\lambda_\beta \in \Lambda_m$ in the following way:

$$G(\lambda_\beta) = \int_0^1 \frac{\lambda_\beta(t)}{\|\lambda_\beta\|} d\bar{M}_n(t).$$

Therefore, we obtain

$$\big\| \hat{\beta} - \beta \big\|_2^2 \le \sqrt{2} \sqrt{L_{max} + 2(L_0 - 1)} \, \big\| \hat{\beta} - \beta \big\|_2 \max_{j=1,\dots,m} \hat{w}_j + 2 \big\| \hat{\beta} - \beta \big\|_2 G(\hat{\beta} - \beta).$$

Let

$$\mathcal{V} = \bigcup_{L=1}^{L_{\max}} V_L = \bigcup_{L=1}^{L_{\max}} \bigcup_{J \subset \{1,\dots,m-1\}, |J|=L} V_{L,J},$$

where $\{V_L : L = 1, \dots, L_{\max}\}$ is the collection of the spaces to which $\hat{\beta}$ may belong and $V_{L,J}$ denotes a space of dimension L containing signals with a support J.

It follows that,

$$\big\| \hat{\beta} - \beta \big\|_2 \le \sqrt{2} \sqrt{L_{\max} + 2(L_0 - 1)} \max_{j=1,\dots,m} \hat{w}_j + 2 \sup_{\lambda \in \mathcal{V}, \|\lambda\|=1} G(\lambda). \tag{2.29}$$

Then by Proposition 4 in Comte et al. (2011), we have for any $z > 0$

$$\mathbb{P}\Big[\sup_{\lambda \in V_{L,J}, \|\lambda\|=1} G(\lambda) \ge \kappa \Big(\sqrt{\frac{\|\lambda_0\|_\infty (L+z)}{n}} + \frac{2\sqrt{m}(L+z)}{\sqrt{L} n} \Big) \Big] \le e^{-z},$$

where $\kappa = 11.8$. Then

$$\sum_{\substack{L=1,\dots,L_{\max}\\ J\subset\{1,\dots,m-1\},|J|=L}} \mathbb{P}\Big[\sup_{\lambda\in V_{L,J},\|\lambda\|=1} G(\lambda) \geq \kappa\Big(\sqrt{\frac{\|\lambda_0\|_\infty(L+z)}{n}} + \frac{2\sqrt{m}(L+z)}{\sqrt{L}n}\Big)\Big]$$
$$\leq \sum_{\substack{L=1,\dots,L_{\max}\\ J\subset\{1,\dots,m-1\},|J|=L}} e^{-z},$$

and

$$\sum_{\substack{L=1,\dots,L_{\max}\\ J\subset\{1,\dots,m-1\},|J|=L}} \mathbb{P}\Big[\sup_{\lambda\in V_{L,J},\|\lambda\|=1} G(\lambda) \geq \kappa\Big(\sqrt{\frac{\|\lambda_0\|_\infty(L+z)}{n}} + \frac{2\sqrt{m}(L+z)}{\sqrt{L}n}\Big)\Big]$$
$$\leq L_{\max} m^{L_{\max}} e^{-z}.$$

Choosing $z = x + L_{\max}\log m$ for $x > 0$, leads to

$$\sum_{\substack{L=1,\dots,L_{\max}\\ J\subset\{1,\dots,m-1\},|J|=L}} \mathbb{P}\Big[\sup_{\lambda\in V_{L,J},\|\lambda\|=1} G(\lambda) \geq \kappa\Big(\sqrt{\frac{\|\lambda_0\|_\infty(L+x+L_{\max}\log m)}{n}}$$
$$+ \frac{2\sqrt{m}(L+x+L_{\max}\log m)}{\sqrt{L}n}\Big)\Big]$$
$$\leq L_{\max} e^{-x}.$$

Plugging this in inequality (2.29), we obtain for any $x > 0$ and with probability larger than $1 - L_{\max}e^{-x}$

$$\big\|\hat\beta - \beta\big\|_2 \leq \sqrt{2}\sqrt{L_{\max} + 2(L_0 - 1)} \max_{j=1,\dots,m} \hat w_j$$
$$+ 2\kappa\sqrt{\frac{\|\lambda_0\|_\infty(x + L_{\max}(1+\log m))}{n}}$$
$$+ 4\kappa\frac{\sqrt{m}(x + L_{\max}(1+\log m))}{n},$$

and the result follows by using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, for all $a, b, c \in \mathbb{R}$. □

2.7 Proof of Theorem 2.4.4

Let us give first the overall structure of the proof, which is inspired from Harchaoui and Lévy-Leduc (2010). In this proof, we repeatedly use the KKT optimality conditions of the optimization problem (2.25), given by Lemma 2.7.1 below. We use also repeatedly deviation arguments of the data-driven weights $\hat w_j$ and a control of the martingale noise, which are provided by Lemma 2.7.2 below. We prove consistency of $\hat\tau_\ell = \frac{\hat j_\ell}{m}$, which is an estimator of the right-hand side boundary $\frac{j_\ell}{m}$ of the interval $I_{j_\ell,m} = (\frac{j_\ell - 1}{m}, \frac{j_\ell}{m}]$, by showing that $\mathbb{P}[A_{n,\ell}] \to 0$ as $n \to \infty$, where $A_{n,\ell} := \{|\hat j_\ell - j_\ell| > \frac{m\varepsilon_n}{2}\}$, for all $\ell \in \{1, \dots, L_0 - 1\}$. We treat separately two cases depending on the positions of j_ℓ and $\hat j_\ell$,. In Case I, we consider $\hat j_\ell < j_\ell$, see Section 2.7.1 and Figure 2.7. In Case II., we

consider $\hat{j}_\ell > j_\ell$, see Section 2.B and Figure 2.8. We decompose even further, using the quantity $\Delta_{j,\min}$ (see Section 2.4), defining the set $C_n = \{\max_{1\le\ell\le L_0-1} |\hat{j}_\ell - j_\ell| < \frac{\Delta_{j,\min}}{2}\}$. We prove that $\mathbb{P}[A_{n,\ell} \cap C_n] \to 0$ and $\mathbb{P}[A_{n,\ell} \cap C_n^{\complement}] \to 0$ as $n \to \infty$ for Case I in Section 2.7.1, and for Case II in Section 2.B.

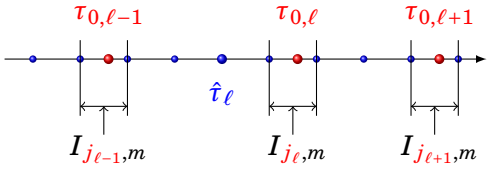


Fig. 2.7 – First case in the proof of Theorem 2.4.4: Case I. $\hat{j}_\ell < j_\ell$.

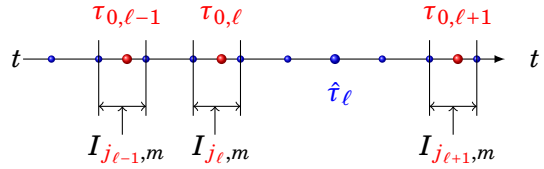


Fig. 2.8 – Second case in the proof of Theorem 2.4.4: Case II. $\hat{j}_\ell > j_\ell$.

Lemma 2.7.1. *Consider the total-variation penalized problems in* (2.21) *and*(2.25). *Let* $\hat{\beta} = [\hat{\beta}_{j,m}]_{1\le j\le m}$ *and* $\hat{\mu} = [\hat{\mu}_{j,m}]_{1\le j\le m}$ *denote the respective solutions. Then, the latter vectors and the approximate change-points sequence estimators* $\hat{j}_1, \dots, \hat{j}_{|\hat{S}|}$ *satisfy for all* $r = 1, \dots, |\hat{S}|$,

$$\sum_{j=\hat{j}_r}^{m} \beta_{0,j,m} - \sum_{j=\hat{j}_r}^{m} \hat{\beta}_{j,m} + \sqrt{m} \sum_{j=\hat{j}_r}^{m} \bar{M}_n(I_{j,m}) = \hat{w}_{\hat{j}_r} \operatorname{sign}(\hat{\mu}_{\hat{j}_r,m}), \tag{2.30}$$

and for all $j \in \{1, \dots, m\}$,

$$\Big| \sum_{q=j}^{m} \beta_{0,q,m} - \sum_{q=j}^{m} \hat{\beta}_{q,m} + \sqrt{m} \sum_{q=j}^{m} \bar{M}_n(I_{q,m}) \Big| \le \hat{w}_j, \tag{2.31}$$

using the convention $\operatorname{sign}(\hat{\mu}_{\hat{j}_r,m}) = +1$, *if* $\hat{\mu}_{\hat{j}_r,m} > 0$ *and* -1 *otherwise. The vectors* $\hat{\beta}$ *and* $\beta_{0,m} = [\beta_{0,j,m}]_{1\le j\le m}$ *have the following additional properties*

$$\begin{cases} \hat{\beta}_{q,m} = \hat{\beta}_{\hat{j}_r-1,m}, & \text{if } \hat{j}_{r-1}+1 \le q \le \hat{j}_r, \text{ for } r = 1, \dots, \hat{L}, \\ \beta_{0,q,m} = \beta_{0,j_\ell-1,m}, & \text{if } j_{\ell-1}+1 \le q \le j_\ell - 1, \text{ for } \ell = 1, \dots, L_0 - 1. \end{cases} \tag{2.32}$$

The proof of Lemma 2.7.1 is given in Section 2.A.3. Let us now state a lemma which allows us to control the martingale noise term.

Lemma 2.7.2. *Given two integers* a *and* b, *such that* $1 \le a < b \le m$, *let* $\bar{M}_n(a;b) := \sum_{q=a}^{b} \bar{M}_n(I_{q,m})$. *Then, for all* $z > 0$ *we have*

$$\mathbb{P}\Big[\big|\bar{M}_n(a;b)\big| \ge z\Big] \le 2\exp\Big\{-\frac{nz^2}{2\int_{\mathbb{1}_{(\frac{a-1}{m},\frac{b}{m}]}} \lambda_0(t)dt + \frac{2}{3}z}\Big\}, \tag{2.33}$$

and for all $\xi > 0$, *the data driven weight* $\hat{w}_a$ *satisfies*

$$\mathbb{P}\Big[\hat{w}_a^2 \ge \frac{m\log m}{n}\Big(\xi - \int_{\mathbb{1}_{(\frac{a-1}{m},1]}} \lambda_0(t)dt\Big)\Big] \le 2\exp\Big\{-\frac{n\xi^2}{2\int_{\mathbb{1}_{(\frac{a-1}{m},1]}} \lambda_0(t)dt + \frac{2}{3}\xi}\Big\}, \tag{2.34}$$

where $\int_I \lambda_0(t)dt = \mathbb{E}[\bar{N}(I)]$ *for any* $I \subset [0,1]$.

The proof of Lemma 2.7.2 is given in Section 2.A.4. Let us now prove Theorem 2.4.4. Recall that the sequence $(\varepsilon_n)_n$ satisfies $m\varepsilon_n \geq 6$, for all $n \geq 1$. An application of the triangle inequality entails that,

$$\mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |\tau_{0,\ell} - \hat{\tau}_\ell| > \varepsilon_n\Big] \leq \mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |\tau_{0,\ell} - \frac{j_\ell}{m}| > \frac{\varepsilon_n}{2}\Big] + \mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |\frac{j_\ell}{m} - \hat{\tau}_\ell| > \frac{\varepsilon_n}{2}\Big].$$

Moreover, the true change-point $\tau_{0,\ell}$ verifies (2.14) which implies that

$$\mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |\tau_{0,\ell} - \hat{\tau}_\ell| > \varepsilon_n\Big] \leq \mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |j_\ell - \hat{j}_\ell| > \frac{m\varepsilon_n}{2}\Big].$$

Due to

$$\mathbb{P}\Big[\max_{1\leq \ell\leq L_0-1} |\hat{j}_\ell - j_\ell| > \frac{m\varepsilon_n}{2}\Big] \leq \sum_{\ell=1}^{L_0-1} \mathbb{P}[|\hat{j}_\ell - j_\ell| > \frac{m\varepsilon_n}{2}],$$

it suffices to prove that for all $\ell = 1,\ldots,L_0-1$, $\mathbb{P}[A_{n,\ell}] \to 0$, as n tending to infinity.

2.7.1 Case I

Due to the fact that $m\varepsilon_n \geq 6$ for all $n \geq 1$, it follows that the event $\{\hat{j}_\ell < j_\ell - 2\}$ a.s.

Step I.1. Prove: $\mathbb{P}[A_{n,\ell} \bigcap C_n] \to 0$, **as** $n \to \infty$.

By the definition of C_n, we have

$$j_{\ell-1} < \hat{j}_\ell < j_{\ell+1}, \text{ for all } \ell = 1,\ldots,L_0-1. \tag{2.35}$$

Applying (2.31) in Lemma 2.7.1 with $j = j_\ell$ and $j = \hat{j}_\ell + 1$, we obtain

$$-(\hat{w}_{j_\ell} + \hat{w}_{\hat{j}_\ell+1}) \leq \sum_{q=\hat{j}_\ell+1}^{j_\ell-1} \mathbf{N}_q - \sum_{q=\hat{j}_\ell+1}^{j_\ell-1} \hat{\beta}_{q,m} \leq \hat{w}_{j_\ell} + \hat{w}_{\hat{j}_\ell+1}.$$

Put $\hat{w}_{a,b} := \hat{w}_a + \hat{w}_b$, for any two integers a and b. Thus

$$\Big|\sum_{q=\hat{j}_\ell+1}^{j_\ell-1} \beta_{0,q,m} - \hat{\beta}_{q,m} + \sqrt{m}\bar{M}_n(I_{q,m})\Big| \leq \hat{w}_{\hat{j}_\ell+1,j_\ell}.$$

Using the property of the vector $\hat{\beta}$ in Lemma 2.7.1, we get

$$\Big|(j_\ell - \hat{j}_\ell - 2)(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_{\ell+1}-1,m}) + \sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)\Big| \leq \hat{w}_{\hat{j}_\ell+1,j_\ell}.$$

Therefore, on $C_n \bigcap \{\hat{j}_\ell < j_\ell - 2\}$, we have

$$\begin{aligned}\Big|&(\hat{j}_\ell - j_\ell - 2)(\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}) \\ &+ (\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}) \\ &+ \sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)\Big| \leq \hat{w}_{\hat{j}_\ell+1,j_\ell}.\end{aligned}$$

Defining the event

$$\begin{aligned}C_{n,\ell} = \Big\{\Big|&(\hat{j}_\ell - j_\ell - 2)(\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}) \\ &+ (\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}) \\ &+ \sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1) + \Big| \leq \hat{w}_{\hat{j}_\ell+1,j_\ell}\Big\},\end{aligned}$$

We observe that $C_{n,\ell}$ occurs with probability one. In addition, we remark that for all $n \geq 1$, $m\varepsilon_n \geq 6$ entails $\frac{m\varepsilon_n}{2} - 2 \geq \frac{m\varepsilon_n}{6}$. Then

$$\left\{|\hat{j}_\ell - j_\ell| > \frac{m\varepsilon_n}{2}\right\} \subset \left\{|\hat{j}_\ell - j_\ell - 2| > \frac{m\varepsilon_n}{2} - 2\right\} \subset \left\{|\hat{j}_\ell - j_\ell - 2| \geq \frac{m\varepsilon_n}{6}\right\}$$

Therefore

$$\begin{aligned}
&\mathbb{P}[A_{n,\ell} \bigcap C_n \bigcap C_{n,\ell}] \\
&\quad \leq \mathbb{P}\left[\left\{\frac{\hat{w}_{\hat{j}_\ell+1,j_\ell}}{|\hat{j}_\ell - j_\ell - 2|} \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\right\} \bigcap \left\{\hat{j}_\ell < j_\ell - 2\right\}\right] \\
&\qquad + \mathbb{P}\left[\left\{|\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\right\} \bigcap C_n\right] \\
&\qquad + \mathbb{P}\left[\left\{\left|\frac{\sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)}{\hat{j}_\ell - j_\ell - 2}\right| \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\right\}\right] \\
&\quad := \mathbb{P}[A_{n,\ell,1}] + \mathbb{P}[A_{n,\ell,2}] + \mathbb{P}[A_{n,\ell,3}].
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\mathbb{P}[A_{n,\ell,1}] &\leq \mathbb{P}\left[\hat{w}_{\hat{j}_\ell+1,j_\ell} \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18}\right] \\
&\leq \mathbb{P}\left[\hat{w}_{\hat{j}_\ell+1} \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{36}\right] \\
&\leq \mathbb{P}\left[\hat{w}^2_{j_{\ell-1}+1} \geq \frac{m^2\varepsilon_n^2\Delta^2_{\beta,\min}}{36^2}\right].
\end{aligned}$$

By (2.16) in Assumption 2.4.3, and (2.34) in Lemma 2.7.2 with $\xi = \frac{nm\varepsilon_n^2\Delta^2_{\beta,\min}}{36^2 \log m} + \mathbb{E}\left[\bar{N}_n\left(\left(\frac{j_{\ell-1}}{m}, 1\right]\right)\right]$,it follows that

$$\mathbb{P}[A_{n,\ell,1}] \quad \leq \quad 2\exp\left\{-\frac{n\xi^2}{2\mathbb{E}\left[\bar{N}_n\left(\left(\frac{j_{\ell-1}}{m}, 1\right]\right)\right] + \frac{2}{3}\xi}\right\} \to 0,$$

as $n \to \infty$. Next, consider the event

$$\begin{aligned}
A_{n,\ell,3} &= \left\{\left|\frac{\sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)}{\hat{j}_\ell - j_\ell - 2}\right| \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\right\} \\
&= \left\{\left|\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)\right| \geq \left|\hat{j}_\ell - j_\ell - 2\right| \frac{\left|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\right|}{3\sqrt{m}}\right\} \\
&\subset \left\{\left|\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)\right| \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18\sqrt{m}}\right\} \bigcap \bigcup_{q=j_{\ell-1}+1}^{j_\ell-3} \{\hat{j}_\ell = q\} \\
&\subset \bigcup_{q=j_{\ell-1}+2}^{j_\ell-2} \left\{\left|\bar{M}_n(q; j_\ell - 1)\right| \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18\sqrt{m}}\right\}.
\end{aligned}$$

Put $\varphi_n = \frac{\sqrt{m}\varepsilon_n \Delta_{\beta,\min}}{18}$. By (2.33) in Lemma 2.7.2, we have

$$\begin{aligned}
\mathbb{P}[A_{n,\ell,3}] &\leq 2 \sum_{q=j_{\ell-1}+2}^{j_\ell-2} \exp\left\{ - \frac{n\varphi_n^2}{2\mathbb{E}\left[\bar{N}_n\left(\left(\frac{q-1}{m}, \frac{j_\ell-1}{m}\right]\right)\right] + \frac{2}{3}\varphi_n} \right\} \\
&\leq 2(j_\ell - j_{\ell-1} - 3)\exp\left\{ - \frac{n\varphi_n^2}{2\mathbb{E}\left[\bar{N}_n\left(\left(\frac{j_{\ell-1}+1}{m}, \frac{j_\ell-1}{m}\right]\right)\right] + \frac{2}{3}\varphi_n} \right\} \\
&\leq 2\exp\left\{ - \frac{n\varphi_n^2}{2\mathbb{E}\left[\bar{N}_n\left(\left(\frac{j_{\ell-1}+1}{m}, \frac{j_\ell-1}{m}\right]\right)\right] + \frac{2}{3}\varphi_n} + \log m \right\}.
\end{aligned}$$

By (2.16) in Assumption 2.4.3 , it implies that $\mathbb{P}[A_{n,\ell,3}]$ goes to zero as $n \to \infty$. We now control $\mathbb{P}[A_{n,\ell,2}]$. Using Lemma 2.7.1 with $j = \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil$ and with $j = j_\ell + 1$, and using the triangle inequality, it follows that

$$\left| \sum_{q=j_\ell+1}^{\lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil - 1} \mathbf{N}_q - \sum_{q=j_\ell+1}^{\lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil - 1} \hat{\beta}_{q,m} \right| \leq \hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil}.$$

Furthermore, on the event $C_n \bigcap \{\hat{j}_\ell < j_\ell - 2\}$, the following inequalities

$$\hat{j}_\ell < j_\ell \leq q \leq \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1 \leq j_{\ell+1} - 1,$$

hold true. Moreover, we note that $\hat{\beta}_{q,m} = \hat{\beta}_{\hat{j}_{\ell+1}-1,m}$ if $j_\ell \leq q \leq \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1 \leq \hat{j}_{\ell+1} - 1$. Consequently, we have

$$\left| (j_{\ell+1} - j_\ell - 2) \frac{(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{\ell+1}-1,m})}{2} + \sqrt{m}\bar{M}_n(j_\ell + 1; \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1) \right| \leq \hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil},$$

which implies that

$$(j_{\ell+1} - j_\ell - 2) \frac{|\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}|}{2} \leq \hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil} + \left| \sqrt{m}\bar{M}_n(j_\ell + 1; \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1) \right|.$$

Therefore, we may upper bound $\mathbb{P}[A_{n,\ell,2}]$ as follows

$$\begin{aligned}
&\mathbb{P}[A_{n,\ell,2}] \\
&= \mathbb{P}\left[\left\{ |\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3} \right\} \bigcap C_n \right] \\
&= \mathbb{P}\left[\left\{ (j_{\ell+1} - j_\ell - 2) \frac{|\hat{\beta}_{\hat{j}_{\ell+1},m} - \beta_{0,j_{\ell+1}-1,m}|}{2} \right. \right. \\
&\qquad \left. \left. \geq (j_{\ell+1} - j_\ell - 2) \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{6} \right\} \bigcap C_n \right] \\
&\leq \mathbb{P}\left[\left\{ \hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil} + \left| \sqrt{m}\bar{M}_n(j_\ell + 1; \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1) \right| \right. \right. \\
&\qquad \left. \left. \geq (j_{\ell+1} - j_\ell - 2) \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{6} \right\} \bigcap C_n \right] \\
&\leq \mathbb{P}\left[\hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil} \geq (j_{\ell+1} - j_\ell - 2) \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{12} \right] \\
&\quad + \mathbb{P}\left[\left| \sqrt{m}\bar{M}_n(j_\ell + 1; \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1) \right| \geq (j_{\ell+1} - j_\ell - 2) \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{12} \right] \\
&\leq \mathbb{P}\left[\hat{w}_{j_\ell+1, \lceil \frac{j_\ell+j_{\ell+1}}{2} \rceil} \geq \frac{(\Delta_{j,\min} - 2)\Delta_{\beta,\min}}{12} \right] \\
&\quad + \mathbb{P}\left[\left| \bar{M}_n(j_\ell + 1; \lceil \frac{j_\ell + j_{\ell+1}}{2} \rceil - 1) \right| \geq \frac{(\Delta_{j,\min} - 2)\Delta_{\beta,\min}}{12\sqrt{m}} \right].
\end{aligned}$$

On the other hand, it is easy to see that (2.13) in Assumption 2.4.1 yields that $\Delta_{j,\min} - 2 \geq \frac{\Delta_{j,\min}}{2} - 2 \geq \frac{\Delta_{j,\min}}{6}$. Thus

$$\begin{aligned}\mathbb{P}[A_{n,\ell,2}] &\leq \mathbb{P}\Big[\hat{w}_{j_\ell+1,\lceil\frac{j_\ell+j_{\ell+1}}{2}\rceil} \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{72}\Big] + \mathbb{P}\Big[\Big|\bar{M}_n(j_\ell+1;\lceil\frac{j_\ell+j_{\ell+1}}{2}\rceil-1)\Big| \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{72\sqrt{m}}\Big] \\ &:= \alpha^{(1)}_{n,\ell,2} + \alpha^{(2)}_{n,\ell,2}.\end{aligned}$$

Using the property of the data-driven weights, we remark that

$$\alpha^{(1)}_{n,\ell,2} \leq \mathbb{P}\Big[\hat{w}^2_{j_\ell+1} \geq \frac{\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{144^2}\Big].$$

By (2.17) in Assumption 2.4.3, and (2.34) in Lemma 2.7.2 with $\xi = \frac{n\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{144^2 m\log m} + \mathbb{E}\big[\bar{N}_n\big((\frac{j_\ell}{m},1]\big)\big]$, it follows that

$$\alpha^{(1)}_{n,\ell,2} \quad \leq \quad 2\exp\Big\{-\frac{n\xi^2}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell}{m},1]\Big)\Big] + \frac{2}{3}\xi}\Big\} \to 0,$$

as $n \to \infty$. Similarly, using (2.17) in Assumption 2.4.3, and (2.33) in Lemma 2.7.2 with $z = \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{72\sqrt{m}}$, it implies that

$$\alpha^{(2)}_{n,\ell,2} \quad \leq \quad 2\exp\Big\{-\frac{nz^2}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell}{m},\frac{\lceil\frac{j_\ell+j_{\ell+1}}{2}\rceil-1}{m}]\Big)\Big] + \frac{2}{3}z}\Big\} \to 0,$$

as $n \to \infty$. Therefore, we conclude that $\mathbb{P}[A_{n,\ell,2}] \to 0$, as $n \to \infty$.

Step I.2. Prove: $\mathbb{P}[A_{n,\ell} \bigcap C_n^{\complement}] \to 0$, **as** $n \to \infty$.

Recall that $C_n^{\complement} = \big\{\max_{1\leq k\leq L_0-1} |\hat{j}_\ell - j_\ell| \geq \frac{\Delta_{j,\min}}{2}\big\}$. We split $\mathbb{P}[A_{n,\ell} \bigcap C_n^{\complement}]$ in three terms as following

$$\mathbb{P}[A_{n,\ell} \bigcap C_n^{\complement}] = \mathbb{P}[A_{n,\ell} \bigcap D_n^{(l)}] + \mathbb{P}[A_{n,\ell} \bigcap D_n^{(m)}] + \mathbb{P}[A_{n,\ell} \bigcap D_n^{(r)}],$$

where

$$\begin{aligned} D_n^{(l)} &:= \{\text{there exists } \ell \in \{1,\dots,L_0-1\} : \hat{j}_\ell \leq j_{\ell-1}\} \bigcap C_n^{\complement} \\ D_n^{(m)} &:= \{\text{for all } \ell \in \{1,\dots,L_0-1\} : j_{\ell-1} < \hat{j}_\ell < j_{\ell+1}\} \bigcap C_n^{\complement}, \\ D_n^{(r)} &:= \{\text{there exists } \ell \in \{1,\dots,L_0-1\} : \hat{j}_\ell \geq j_{\ell+1}\} \bigcap C_n^{\complement}. \end{aligned}$$

Let us first focus on $\mathbb{P}[A_{n,\ell} \bigcap D_n^{(m)}]$, see Figure 2.9. Observe that

$$\mathbb{P}[A_{n,\ell} \bigcap D_n^{(m)}] = \mathbb{P}\Big[A_{n,\ell} \bigcap \{\hat{j}_{\ell+1} - j_\ell \geq \frac{\Delta_{j,\min}}{2}\} \bigcap D_n^{(m)}\Big] + \mathbb{P}[A_{n,\ell} \bigcap \{\hat{j}_{\ell+1} - j_\ell < \frac{\Delta_{j,\min}}{2}\} \bigcap D_n^{(m)}].$$

The fact that $0 \leq \hat{j}_{\ell+1} - j_\ell < \frac{\Delta_{j,\min}}{2}$ yields $j_{\ell+1} - \hat{j}_{\ell+1} \geq \frac{\Delta_{j,\min}}{2}$. Then, it is easy to see that $j_{\ell+1} - \hat{j}_{\ell+1} = (j_{\ell+1} - j_\ell) - (\hat{j}_{\ell+1} - j_\ell) \geq \Delta_{j,\min} - \frac{\Delta_{j,\min}}{2} \geq \frac{\Delta_{j,\min}}{2}$.
Hence

$$\mathbb{P}[A_{n,\ell} \bigcap D_n^{(m)}] \leq \mathbb{P}\Big[A_{n,\ell} \bigcap \{\hat{j}_{\ell+1} - j_\ell \geq \frac{\Delta_{j,\min}}{2}\} \bigcap D_n^{(m)}\Big] + \mathbb{P}\Big[A_{n,\ell} \bigcap \{j_{\ell+1} - \hat{j}_{\ell+1} \geq \frac{\Delta_{j,\min}}{2}\} \bigcap D_n^{(m)}\Big].$$

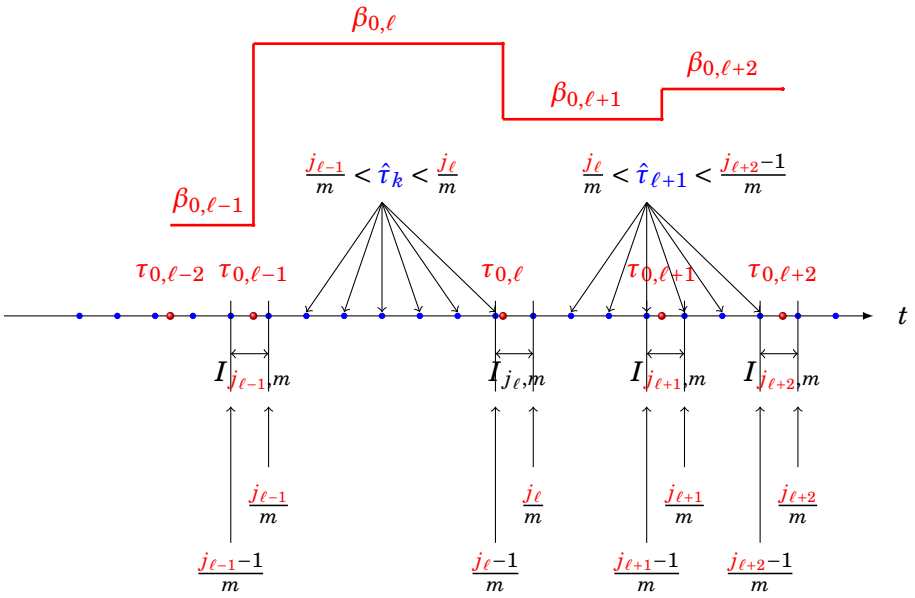


Fig. 2.9 – A zoom into the Case I.

Moreover, we note that

$$A_{n,\ell} \bigcap \Big\{ j_{\ell+1} - \hat{j}_{\ell+1} \geq \frac{\Delta_{j,\min}}{2} \Big\} \bigcap D_n^{(m)} \subset \bigcup_{r=\ell+1}^{L_0-2} \Big\{ j_r - \hat{j}_r \geq \frac{\Delta_{j,\min}}{2} \Big\} \bigcap \Big\{ \hat{j}_{r+1} - j_r \geq \frac{\Delta_{j,\min}}{2} \Big\} \bigcap D_n^{(m)}.$$

Thus, we have

$$\mathbb{P}[A_{n,\ell} \bigcap D_n^{(m)}] \leq \mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)}] + \sum_{s=\ell+1}^{L_0-2} \mathbb{P}[C_{s,s} \bigcap B_{s+1,s} \bigcap D_n^{(m)}], \tag{2.36}$$

where

$$\begin{cases} B_{p,q} = \{(\hat{j}_p - j_q) \geq \frac{\Delta_{j,\min}}{2}\}, \\ \text{with the convention } B_{L_0,L_0-1} = \{m - j_{L_0-1} \geq \frac{\Delta_{j,\min}}{2}\}, \\ C_{p,q} = \{(j_p - \hat{j}_q) \geq \frac{\Delta_{j,\min}}{2}\}. \end{cases}$$

Let us now prove that the first term in the right hand side of (2.36) goes to zero as n tends to infinity , the arguments for the other terms being similar. Using (2.31) in Lemma 2.7.1 with $j = j_\ell$ and $j = \hat{j}_\ell + 1$, on the one hand and (2.31) in Lemma 2.7.1 with $j = j_\ell + 1$ and $j = \hat{j}_{\ell+1}$ on the other hand, we obtain, respectively

$$|\hat{j}_\ell - j_\ell - 2||\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}| \leq \hat{w}_{\hat{j}_{\ell+1},j_\ell} + \big|\sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)\big|, \tag{2.37}$$

and

$$|\hat{j}_{\ell+1} - j_\ell - 2||\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \leq \hat{w}_{j_\ell+1,\hat{j}_{\ell+1}} + |\sqrt{m}\bar{M}_n(j_\ell + 1; \hat{j}_{\ell+1} - 1)|. \tag{2.38}$$

Besides, we have

$$
\begin{aligned}
&|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}| \\
&\quad = |(\hat\beta_{\hat j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}) - (\hat\beta_{\hat j_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1})| \\
&\quad \le |\hat\beta_{\hat j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}| + |\hat\beta_{\hat j_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \\
&\quad \le \frac{\hat w_{\hat j_\ell+1,j_\ell}}{|\hat j_\ell - j_\ell - 2|} + \frac{\sqrt{m}\bar M_n(\hat j_\ell+1; j_\ell-1)|}{|\hat j_\ell - j_\ell - 2|} \\
&\qquad\qquad + \frac{\hat w_{j_\ell+1,\hat j_{\ell+1}}}{|\hat j_{\ell+1} - j_\ell - 2|} + \frac{|\sqrt{m}\bar M_n(j_\ell+1; \hat j_{\ell+1}-1)|}{|\hat j_{\ell+1} - j_\ell - 2|} \\
&\quad \le \frac{\hat w_{\hat j_\ell+1,j_\ell}}{\frac{m\varepsilon_n}{6}} + \frac{\sqrt{m}\bar M_n(\hat j_\ell+1; j_\ell-1)|}{|\hat j_\ell - j_\ell - 2|} \\
&\qquad\qquad + \frac{\hat w_{j_\ell+1,\hat j_{\ell+1}}}{\frac{\Delta_{j,\min}}{6}} + \frac{|\sqrt{m}\bar M_n(j_\ell+1; \hat j_{\ell+1}-1)|}{|\hat j_{\ell+1} - j_\ell - 2|}.
\end{aligned}
$$

Define the event $E_{n,\ell}$ by

$$
\begin{aligned}
E_{n,\ell} = \Bigg\{ |\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}| \le{}& \frac{\hat w_{\hat j_\ell+1,j_\ell}}{\frac{m\varepsilon_n}{6}} + \frac{\hat w_{j_\ell+1,\hat j_{\ell+1}}}{\frac{\Delta_{j,\min}}{6}} \\
&+ \left|\frac{\sqrt{m}\bar M_n(\hat j_\ell+1; j_\ell-1)}{\hat j_\ell - j_\ell - 2}\right| \\
&+ \left|\frac{\sqrt{m}\bar M_n(j_\ell+1; \hat j_{\ell+1}-1)}{\hat j_{\ell+1} - j_\ell - 2}\right| \Bigg\}.
\end{aligned}
$$

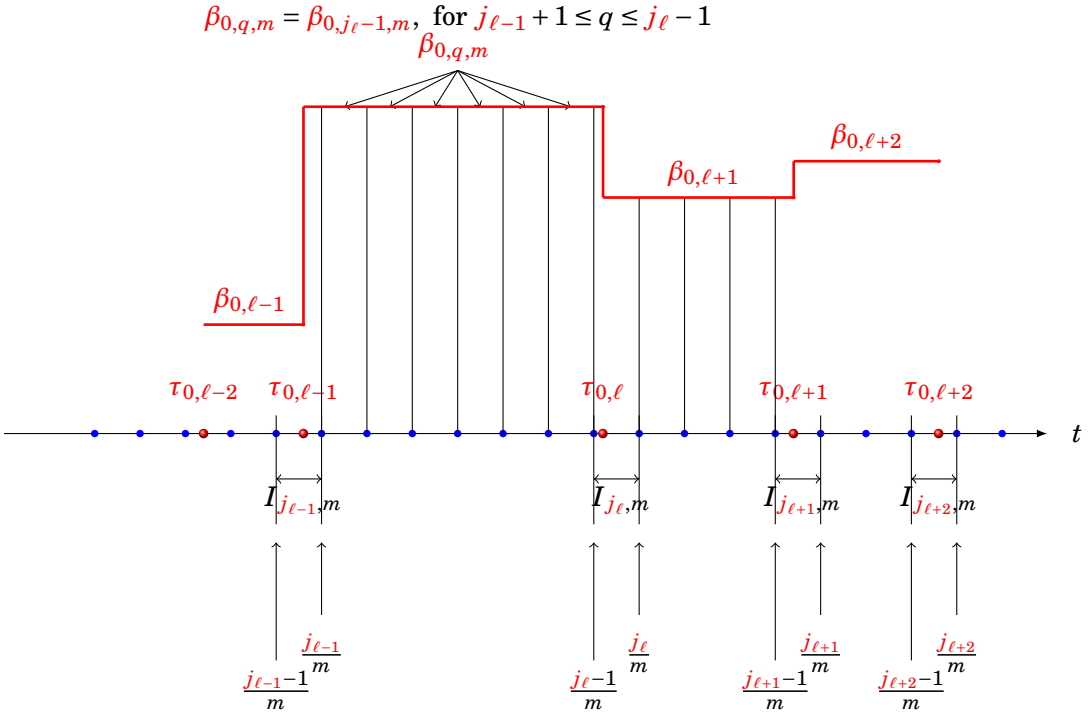


Fig. 2.10 – A zoom of $\beta_{0,q,m}$, the coefficients of the projection function $\lambda_{0,m}$ in Case I.

We observe that $E_{n,\ell}$ occurs with probability one, see Figure 2.10. Therefore, we obtain

$$
\begin{aligned}
&\mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)}] \\
&\quad \leq \mathbb{P}\Big[E_{n,\ell} \bigcap \{(j_\ell - \hat{j}_\ell) > \frac{m\varepsilon_n}{2}\} \bigcap \{(\hat{j}_{\ell+1} - j_\ell) \geq \frac{\Delta_{j,\min}}{2}\}\Big] \\
&\quad \leq \mathbb{P}\Big[\hat{w}_{\hat{j}_\ell+1,j_\ell} \geq \frac{m\varepsilon_n |\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{24}\Big] \\
&\qquad + \mathbb{P}\Big[\hat{w}_{j_\ell+1,\hat{j}_{\ell+1}} \geq \frac{\Delta_{j,\min} |\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{24}\Big] \\
&\qquad + \mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)}{j_\ell - \hat{j}_\ell - 2}\Big| \\
&\qquad\qquad \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\Big\} \bigcap \Big\{j_\ell - \hat{j}_\ell - 2 \geq \frac{m\varepsilon_n}{6}\Big\}\Big] \\
&\qquad + \mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(j_\ell + 1; \hat{j}_{\ell+1} - 1)}{\hat{j}_{\ell+1} - j_\ell - 2}\Big| \\
&\qquad\qquad \geq \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\Big\} \bigcap \Big\{\hat{j}_{\ell+1} - j_\ell - 2 \geq \frac{\Delta_{j,\min}}{6}\Big\}\Big]. \\
&\quad := \theta_{n,\ell,1} + \theta_{n,\ell,2} + \theta_{n,\ell,3} + \theta_{n,\ell,4}
\end{aligned}
$$

We note

$$
\theta_{n,\ell,1} \leq \mathbb{P}\Big[\hat{w}_{\hat{j}_\ell+1} \geq \frac{m\varepsilon_n \Delta_{\beta,\min}}{48}\Big] \leq \mathbb{P}\Big[\hat{w}^2_{j_{\ell-1}+1} \geq \frac{m^2\varepsilon_n^2 \Delta^2_{\beta,\min}}{48^2}\Big].
$$

Using (2.16) in Assumption 2.4.3, and (2.34) in Lemma 2.7.2 with $\xi = \frac{nm\varepsilon_n^2\Delta^2_{\beta,\min}}{48^2 \log m} + \mathbb{E}\Big[\bar{N}_n\Big(\big(\frac{j_{\ell-1}}{m}, 1\big]\Big)\Big]$, we get

$$
\theta_{n,\ell,1} \quad \leq \quad 2\exp\Big\{-\frac{n\xi^2}{2\mathbb{E}\Big[\bar{N}_n\Big(\big(\frac{j_{\ell-1}}{m}, 1\big]\Big)\Big] + \frac{2}{3}\xi}\Big\} \to 0,
$$

as $n \to \infty$. Analogously,

$$
\theta_{n,\ell,2} \leq \mathbb{P}\Big[\hat{w}_{\hat{j}_\ell+1} \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{48}\Big] \leq \mathbb{P}\Big[\hat{w}^2_{j_\ell+1} \geq \frac{\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{48^2}\Big].
$$

Using (2.17) in Assumption 2.4.3, and (2.34) in Lemma 2.7.2, with $\xi = \frac{n\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{48^2 m \log m} + \mathbb{E}[\bar{N}_n((\frac{j_\ell}{m}, 1])]$, we have

$$
\theta_{n,\ell,2} \quad \leq \quad 2\exp\Big\{-\frac{n\xi^2}{2\mathbb{E}\Big[\bar{N}_n\Big(\big(\frac{j_\ell}{m}, 1\big]\Big)\Big] + \frac{2}{3}\xi}\Big\} \to 0,
$$

as $n \to \infty$. Furthermore, using (2.33) in Lemma 2.7.2, we have

$$
\begin{aligned}
\theta_{n,\ell,3} \quad &\leq \quad \mathbb{P}\Big[\,|\bar{M}_n(\hat{j}_\ell + 1; j_\ell - 1)| \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{24\sqrt{m}}\Big] \\
&\leq \quad 2\exp\Big\{-\frac{n\psi_n^2}{2\mathbb{E}\Big[\bar{N}_n\Big(\big(\frac{j_{\ell-1}+1}{m}, \frac{j_\ell-1}{m}\big]\Big)\Big] + \frac{2}{3}\psi_n} + \log m\Big\},
\end{aligned}
$$

where $\psi_n = \frac{\sqrt{m}\varepsilon_n \Delta_{\beta,\min}}{24}$. By (2.16) in Assumption 2.4.3, we get that $\theta_{n,\ell,3} \to 0$, as $n \to \infty$. Similarly, using (2.33) in Lemma 2.7.2, we have

$$\begin{aligned}\theta_{n,\ell,4} &\leq \mathbb{P}\Big[\big|\bar{M}_n(j_\ell + 1; \hat{j}_{\ell+1} - 1)\big| \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{24\sqrt{m}}\Big] \\ &\leq 2\exp\Big\{-\frac{n\delta_n^2}{2\mathbb{E}\Big[\bar{N}_n\Big(\big(\frac{j_\ell}{m}, \frac{j_{\ell+2}-2}{m}\big]\Big)\Big] + \frac{2}{3}\delta_n} + \log m\Big\},\end{aligned}$$

where $\delta_n = \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{24\sqrt{m}}$. By (2.17) in Assumption 2.4.3, we get that $\theta_{n,\ell,4} \to 0, \text{as } n \to \infty$. Consequently, we obtain $\mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)}] \to 0$ as $n \to \infty$. Now, we have $\mathbb{P}[A_{n,\ell} \bigcap {D_n}^{(l)}] \leq \mathbb{P}[{D_n}^{(l)}]$, and

$$\begin{aligned}\mathbb{P}[{D_n}^{(l)}] &= \mathbb{P}\Big[\{\exists\, \ell \in \{1,\ldots,L_0-1\} : \hat{j}_\ell \leq j_{\ell-1}\} \bigcap C_n^{\complement}\Big] \\ &= \mathbb{P}\Big[\Big\{\bigcup_{\ell=1}^{L_0-1} \max\{1 \leq q \leq L_0 - 1 : \hat{j_q} \leq j_{q-1}\} = \ell\Big\} \bigcap C_n^{\complement}\Big] \\ &= \sum_{\ell=1}^{L_0-1} \mathbb{P}\Big[\{\max\{1 \leq q \leq L_0 - 1 : \hat{j_q} \leq j_{q-1}\} = \ell\} \bigcap C_n^{\complement}\Big].\end{aligned}$$

We note that on the event $\{\max\{1 \leq q \leq L_0 - 1; \hat{j_q} \leq j_{q-1}\} = \ell\}$, it is clear to see that $\hat{j}_\ell \leq j_{\ell-1}$ and $\hat{j}_{q+1} > j_q$ for all $q = \ell, \ldots, L_0 - 1$. Then, it follows that

$$\mathbb{P}[{D_n}^{(l)}] \leq \sum_{\ell=1}^{L_0-1} 2^{\ell-1}\mathbb{P}\Big[\bigcap_{q\geq\ell}^{L_0-1} \{\hat{j}_\ell \leq j_{\ell-1}\} \bigcap \{\hat{j}_{q+1} > j_q\}\Big].$$

In addition, we note that

$$\begin{aligned}&\bigcap_{q\geq\ell}^{L_0-1} \{\hat{j}_\ell \leq j_{\ell-1}\} \bigcap \{\hat{j}_{q+1} > j_q\} \\ &\subset \{j_\ell \leq \hat{j_\ell}\} \bigcap \Big(\{\hat{j}_{\ell+1} > \frac{j_\ell + j_{\ell+1}}{2}\} \bigcup \{\hat{j}_{\ell+1} < \frac{j_\ell + j_{\ell+1}}{2}\}\Big) \\ &\quad \bigcap \Big(\{\hat{j}_{\ell+2} > \frac{j_{\ell+1} + j_{\ell+2}}{2}\} \bigcup \{\hat{j}_{\ell+2} < \frac{j_{\ell+2} + j_{\ell+1}}{2}\}\Big) \\ &\quad \bigcap \ldots \bigcap \Big(\{\hat{j}_{L_0-1} > \frac{j_{L_0-2} + j_{L_0+1}}{2}\} \bigcup \{\hat{j}_{L_0-1} < \frac{j_{L_0-2} + j_{L_0+1}}{2}\}\Big) \\ &\quad \bigcap \Big(\{\hat{j}_{L_0} > \frac{j_{L_0-1} + j_{L_0}}{2}\} \bigcup \{\hat{j}_{L_0} < \frac{j_{L_0-1} + j_{L_0}}{2}\}\Big) \\ &\subset \{j_\ell - \hat{j_\ell} > \frac{\Delta_{j,\min}}{2}\} \bigcap \Big(\{\hat{j}_{\ell+1} - j_\ell > \frac{\Delta_{j,\min}}{2}\} \bigcup \{j_{\ell+1} - \hat{j}_{\ell+1} > \frac{\Delta_{j,\min}}{2}\}\Big) \\ &\quad \bigcap \Big(\{\hat{j}_{\ell+2} - j_{\ell+1} > \frac{\Delta_{j,\min}}{2}\} \bigcup \{j_{\ell+2} - \hat{j}_{\ell+2} > \frac{\Delta_{j,\min}}{2}\}\Big) \\ &\quad \bigcap \ldots \bigcap \Big(\{\hat{j}_{L_0-1} - j_{L_0-2} > \frac{\Delta_{j,\min}}{2}\} \bigcup \{j_{L_0-2} - \hat{j}_{L_0-2} > \frac{\Delta_{j,\min}}{2}\}\Big) \\ &\quad \bigcap \Big(\{\hat{j}_{L_0} - j_{L_0-1} > \frac{\Delta_{j,\min}}{2}\} \bigcup \{j_{L_0-1} - \hat{j}_{L_0-1} > \frac{\Delta_{j,\min}}{2}\}\Big) \\ &\subset \bigcup_{q=\ell}^{L_0-2} \Big(\{j_q - \hat{j_q} > \frac{\Delta_{j,\min}}{2}\} \bigcap \{\hat{j}_{q+1} - j_q > \frac{\Delta_{j,\min}}{2}\}\Big) \bigcup \{j_{L_0-1} - \hat{j}_{L_0-1} > \frac{\Delta_{j,\min}}{2}\}.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{P}[D_n{}^{(l)}] \leq 2^{L_0-2}\sum_{\ell=1}^{L_0-2}\sum_{q=\ell}^{L_0-2}\mathbb{P}\Big[\{(j_q-\hat{j}_q)>\frac{\Delta_{j,\min}}{2}\}\bigcap\{\hat{j}_{q+1}-j_q>\frac{\Delta_{j,\min}}{2}\}\Big] \\ +2^{L_0-2}\mathbb{P}\Big[\{j_{L_0-1}-\hat{j}_{L_0-1}>\frac{\Delta_{j,\min}}{2}\}\Big].\end{aligned} \tag{2.39}$$

Consider the first term of the sum in the right-hand side of (2.39). Using (2.37) and (2.38) with $\ell = q$, we obtain

$$\begin{aligned}&\mathbb{P}\Big[\Big\{j_q-\hat{j}_q>\frac{\Delta_{j,\min}}{2}\Big\}\bigcap\Big\{\hat{j}_{q+1}-j_q>\frac{\Delta_{j,\min}}{2}\Big\}\Big] \\ &\quad\leq \mathbb{P}\Big[\frac{\hat{w}_{\hat{j}_q+1,j_q}}{\frac{\Delta_{j,\min}}{6}}\geq\frac{|\beta_{0,j_{q+1}-1,m}-\beta_{0,j_q-1,m}|}{4}\Big] \\ &\qquad+\mathbb{P}\Big[\frac{\hat{w}_{j_q+1,\hat{j}_{q+1}}}{\frac{\Delta_{j,\min}}{6}}\geq\frac{|\beta_{0,j_{q+1}-1,m}-\beta_{0,j_q-1,m}|}{4}\Big] \\ &\qquad+\mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1;j_q-1)}{j_q-\hat{j}_q-2}\Big| \\ &\qquad\qquad\geq\frac{|\beta_{0,j_{q+1}-1,m}-\beta_{0,j_q-1,m}|}{4}\Big\}\bigcap\Big\{j_q-\hat{j}_q\geq\frac{\Delta_{j,\min}}{2}\Big\}\Big] \\ &\qquad+\mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(j_q+1;\hat{j}_{q+1}-1)}{\hat{j}_{q+1}-j_q-2}\Big| \\ &\qquad\qquad\geq\frac{|\beta_{0,j_{q+1}-1,m}-\beta_{0,j_q-1,m}|}{4}\Big\}\bigcap\Big\{\hat{j}_{q+1}-j_q\geq\frac{\Delta_{j,\min}}{2}\Big\}\Big]. \\ &:=\theta_{n,q,1}+\theta_{n,q,2}+\theta_{n,q,3}+\theta_{n,q,4}.\end{aligned}$$

By (2.33)-(2.34) in Lemma 2.7.2, and (2.16)-(2.17) in Assumption 2.4.3, we show that for $s = 1,\ldots,4$, $\theta_{n,q,s}\to 0$, as n tending to infinity. Then

$$\mathbb{P}\Big[\Big\{j_q-\hat{j}_q>\frac{\Delta_{j,\min}}{2}\Big\}\bigcap\Big\{\hat{j}_{q+1}-j_q>\frac{\Delta_{j,\min}}{2}\Big\}\Big]\to 0.$$

Let us now consider the last term in the right hand of (2.39). Using the observations (2.37) and (2.38) with $\ell = L_0-1$ leads to

$$\begin{aligned}&\mathbb{P}\Big[\Big\{j_{L_0-1}-\hat{j}_{L_0-1}>\frac{\Delta_{j,\min}}{2}\Big\}\Big] \\ &\quad\leq \mathbb{P}\Big[\frac{\hat{w}_{\hat{j}_{L_0-1}+1,j_{L_0-1}}}{\frac{m\varepsilon_n}{6}}\geq\frac{|\beta_{0,j_{L_0}-1,m}-\beta_{0,j_{L_0-1}-1,m}|}{4}\Big] \\ &\qquad+\mathbb{P}\Big[\frac{\hat{w}_{j_{L_0-1}+1,m}}{\frac{\Delta_{j,\min}}{6}}\geq\frac{|\beta_{0,j_{L_0}-1,m}-\beta_{0,j_{L_0-1}-1,m}|}{4}\Big] \\ &\qquad+\mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_{L_0-1}+1;j_{L_0-1}-1)}{j_{L_0-1}-\hat{j}_{L_0-1}-2}\Big| \\ &\qquad\qquad\geq\frac{|\beta_{0,j_{L_0}-1,m}-\beta_{0,j_{L_0-1}-1,m}|}{4}\Big\}\bigcap\Big\{j_{L_0-1}-\hat{j}_{L_0-1}\geq\frac{\Delta_{j,\min}}{2}\Big\}\Big] \\ &\qquad+\mathbb{P}\Big[\Big\{\Big|\frac{\sqrt{m}\bar{M}_n(j_{L_0-1}+1;m-1)}{m-j_{L_0-1}-2}\Big|\geq\frac{|\beta_{0,j_{L_0}-1,m}-\beta_{0,j_{L_0-1}-1,m}|}{4}\Big\}\Big] \\ &:=\theta_{n,L_0-1,1}+\theta_{n,L_0-1,2}+\theta_{n,L_0-1,3}+\theta_{n,L_0-1,4}.\end{aligned}$$

By (2.33)-(2.34) in Lemma 2.7.2, and (2.16)-(2.17) in Assumption 2.4.3, we show that for $s = 1,\dots,4$, we obtain $\theta_{n,L_0-1,s} \to 0$, as $n \to \infty$. Then

$$\mathbb{P}\Big[\Big\{j_{L_0-1} - \hat{j}_{L_0-1} > \frac{\Delta_{j,\min}}{2}\Big\} \bigcap \Big\{m - j_{L_0-1} > \frac{\Delta_{j,\min}}{2}\Big\}\Big] \to 0.$$

This implies that $\mathbb{P}[D_n{}^{(l)}] \to 0$, as $n \to \infty$. Similarly, we prove that $\mathbb{P}[D_n{}^{(r)}] \to 0$, as $n \to \infty$ which yields that $\mathbb{P}[A_{n,\ell} \bigcap C_n^{\complement}] \to 0$, as $n \to \infty$. This concludes the proof of Theorem 2.4.4, up to the case $\{\hat{j}_\ell > j_\ell\}$ for a fixed $\ell \in \{1,\dots,L_0-1\}$ which is given in Section 2.B.

2.8 Proof of Theorem 2.4.5

This proof is based on the same arguments in the proof of Theorem 2.4.4. Let $\mathcal{T}_0^{\text{approx}} = \{\frac{j_1}{m},\dots,\frac{j_{L_0-1}}{m}\}$ be the set of the true approximate change-points. First, we note that

$$\mathbb{P}\Big[\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0) > \varepsilon_n\Big] \le \mathbb{P}\Big[\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big] + \mathbb{P}\Big[\mathscr{E}(\mathcal{T}_0^{\text{approx}})\|\mathcal{T}_0) > \varepsilon_n\Big].$$

Obviously, since $m\varepsilon_n \ge 6$, we have $\mathbb{P}\Big[\mathscr{E}(\mathcal{T}_0^{\text{approx}})\|\mathcal{T}_0) > \varepsilon_n\Big] = 0$. It is clear to remark that the inequality $\hat{L} \le m$ holds true. In order to prove that

$$\mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{\hat{L} \ge L_0 - 1\Big\}\Big] \to 0,$$

as $n \to \infty$, it is enough to prove that

$$\mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{L_0 - 1 \le \hat{L} \le m\Big\}\Big] \to 0,$$

as $n \to \infty$. We have that

$$\begin{aligned}
&\mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{L_0 - 1 \le \hat{L} \le m\Big\}\Big] \\
&\le \mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{\mathbb{1}_{\hat{L}=L_0-1}\Big\}\Big] + \mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{\mathbb{1}_{\hat{L}>L_0-1}\Big\}\Big] \\
&\le \mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{\mathbb{1}_{\hat{L}=L_0-1}\Big\}\Big] + \sum_{L=L_0}^{m} \mathbb{P}\Big[\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big] \\
&\le \mathbb{P}\Big[\Big\{\mathscr{E}(\hat{\mathcal{T}}\|\mathcal{T}_0^{\text{approx}}) > \varepsilon_n\Big\} \bigcap \Big\{\mathbb{1}_{\hat{L}=L_0-1}\Big\}\Big] \\
&\qquad + \sum_{L=L_0}^{m} \sum_{\ell=1}^{L_0-1} \mathbb{P}\Big[\forall q \in \{1,\dots,L\}, |\frac{\hat{j}_q}{m} - \frac{j_\ell}{m}| > \varepsilon_n\Big]
\end{aligned} \tag{2.40}$$

The first term of the right-hand side of (2.40) tends to zero as n tends to infinity since it is upper bounded by $\mathbb{P}\Big[\max_{1\le\ell\le L_0-1} |\hat{j}_\ell - j_\ell| > m\varepsilon_n\Big]$ which tends to zero by the proof of Theorem 2.4.4. Let us now focus on the second term on the right-hand side of (2.40). Note that

$$\sum_{L=L_0}^{m} \sum_{\ell=1}^{L_0-1} \mathbb{P}\Big[\forall\, 1 \le q \le L, |\hat{j}_q - j_\ell| > m\varepsilon_n\Big] := \sum_{L=L_0}^{m} \sum_{\ell=1}^{L_0-1} \mathbb{P}[R_{n,\ell,1}] + \mathbb{P}[R_{n,\ell,2}] + \mathbb{P}[R_{n,\ell,3}],$$

where

$$R_{n,\ell,1} := \Big\{\forall\, 1 \le q \le L : |\hat{j}_q - j_\ell| > m\varepsilon_n \text{ and } \hat{j}_q < j_\ell\Big\}$$
$$R_{n,\ell,2} := \Big\{\forall\, 1 \le q \le L : |\hat{j}_q - j_\ell| > m\varepsilon_n \text{ and } \hat{j}_q > j_\ell\Big\}$$
$$R_{n,\ell,3} := \Big\{\exists\, 1 \le q \le L-1 : \{|\hat{j}_q - j_\ell| > m\varepsilon_n\}, \{|\hat{j}_{q+1} - j_\ell| > m\varepsilon_n\}, \text{ and } \{\hat{j}_q < j_\ell < \hat{j}_{q+1}\}\Big\}.$$

Note that

$$\mathbb{P}[R_{n,\ell,1}] = \mathbb{P}\Big[R_{n,\ell,1} \bigcap \{\hat{j}_L > j_{\ell-1}\}\Big] + \mathbb{P}\Big[R_{n,\ell,1} \bigcap \{\hat{j}_L \le j_{\ell-1}\}\Big].$$

By applying (2.31) in Lemma 2.7.1 with $j = j_\ell$ and with $j = \hat{j}_L + 1$ in the case where $\hat{j}_L > j_{\ell-1}$, it follows that, with probability one,

$$\Big|(j_\ell - \hat{j}_L - 2)\big((\beta_{0,j_\ell - 1,m} - \beta_{0,j_{\ell+1}-1,m}) + (\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{L+1}-1,m}) + \sqrt{m}\bar{M}_n(\hat{j}_L + 1; j_\ell - 1)\Big| \le \hat{w}_{\hat{j}_L+1,j_\ell}.$$

Thus

$$\begin{aligned}
&\mathbb{P}\Big[R_{n,\ell,1} \bigcap \{\hat{j}_L > j_{\ell-1}\}\Big] \\
&\le \mathbb{P}\Big[\Big\{\frac{\hat{w}_{\hat{j}_L+1,j_\ell}}{m\varepsilon_n - 2} \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big\} \bigcap \{\hat{j}_L > j_{\ell-1}\}\Big] \\
&\quad + \mathbb{P}\Big[\Big\{|\hat{\beta}_{\hat{j}_{L+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big\}\Big] \\
&\quad + \mathbb{P}\Big[\Big\{\Big|\frac{\bar{M}_n(\hat{j}_L + 1; j_\ell - 1)}{j_\ell - \hat{j}_L - 2}\Big| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3\sqrt{m}}\Big\} \bigcap \Big\{|j_\ell - \hat{j}_L| \ge m\varepsilon_n\Big\}\Big] \\
&:= \mathbb{P}[R^{(1)}_{n,\ell,1}] + \mathbb{P}[R^{(2)}_{n,\ell,1}] + \mathbb{P}[R^{(3)}_{n,\ell,1}].
\end{aligned}$$

Using (2.16) in Assumption 2.4.3, and (2.33)-(2.34) in Lemma 2.7.2 with $\xi = \frac{nm\varepsilon_n^2\Delta^2_{\beta,\min}}{36^2\log m} + \mathbb{E}[\bar{N}_n((\frac{j_{\ell-1}}{m}, 1])]$, we prove that $\sum_{L=L_0}^{m}\sum_{\ell=1}^{L_0-1}\mathbb{P}[R^{(3)}_{n,\ell,1}] \to 0, as\, n \to \infty$. Let us now consider to $\mathbb{P}[R^{(2)}_{n,\ell,2}]$. Using (2.31) in Lemma 2.7.1 with $j = j_\ell + 1$ and with $j = j_{\ell+1}$, we get

$$(j_{\ell+1} - j_\ell - 2)|\hat{\beta}_{\hat{j}_{L+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \le \hat{w}_{j_\ell+1,j_{\ell+1}} + |\sqrt{m}\bar{M}_n(j_\ell + 1; j_{\ell+1} - 1)|.$$

Therefore, we may upper bound $\mathbb{P}[R^{(2)}_{n,\ell,2}]$ as follows:

$$\begin{aligned}
&\mathbb{P}\Big[|\hat{\beta}_{\hat{j}_{L+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big] \\
&\le \mathbb{P}\Big[\hat{w}_{j_\ell+1,j_{\ell+1}} \ge (j_{\ell+1} - j_\ell - 2)\frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{6}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\bar{M}_n(j_\ell + 1; j_{\ell+1} - 1)}{j_{\ell+1} - j_\ell - 2}\Big| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{6\sqrt{m}}\Big].
\end{aligned}$$

By using Lemma 2.7.1, and (2.16)-(2.17) in Assumption 2.4.3, we conclude that $\sum_{L=L_0}^{m}\sum_{\ell=1}^{L_0-1}\mathbb{P}(R^{(2)}_{n,\ell,1}) \to 0, as\, n \to \infty$. Analogously, it can be shown that $\sum_{L=L_0}^{m}\sum_{\ell=1}^{L_0-1}\mathbb{P}\Big[R_{n,\ell,1} \bigcap \{\hat{j}_L \le j_{\ell-1}\}\Big] \to 0, as\, n \to \infty$. Moreover, we prove, similarly, that

$\sum_{L=L_0}^{m}\sum_{\ell=1}^{L_0-1}\mathbb{P}[R_{n,\ell,2}] \to 0, as\, n \to \infty$. Let us now focus on $\sum_{L=L_0}^{m}\sum_{\ell=1}^{L_0-1}\mathbb{P}[R_{n,\ell,3}]$. Note that $\mathbb{P}[R_{n,\ell,3}]$ can be split in four terms as follows:

$$\mathbb{P}[R_{n,\ell,3}] = \mathbb{P}[R^{(1)}_{n,\ell,3}] + \mathbb{P}[R^{(2)}_{n,\ell,3}] + \mathbb{P}[R^{(3)}_{n,\ell,3}] + \mathbb{P}[R^{(4)}_{n,\ell,3}],$$

where

$$\begin{aligned}
R^{(1)}_{n,\ell,3} &:= R_{n,\ell,3} \bigcap \left\{ j_{\ell-1} < \hat{j}_q < \hat{j}_{q+1} < j_{\ell+1} \right\} \\
R^{(2)}_{n,\ell,3} &:= R_{n,\ell,3} \bigcap \left\{ j_{\ell-1} < \hat{j}_q < j_{\ell+1}, \hat{j}_{q+1} \geq j_{\ell+1} \right\} \\
R^{(3)}_{n,\ell,3} &:= R_{n,\ell,3} \bigcap \left\{ \hat{j}_q \leq j_{\ell-1}, j_{\ell-1} < \hat{j}_{q+1} < j_{\ell+1} \right\} \\
R^{(4)}_{n,\ell,3} &:= R_{n,\ell,3} \bigcap \left\{ \hat{j}_q \leq j_{\ell-1}, j_{\ell+1} \leq \hat{j}_{q+1} \right\}.
\end{aligned}$$

We have to use Lemma 2.7.1 twice. For $\mathbb{P}[R^{(1)}_{n,\ell,3}]$, we first use (2.31) in Lemma 2.7.1

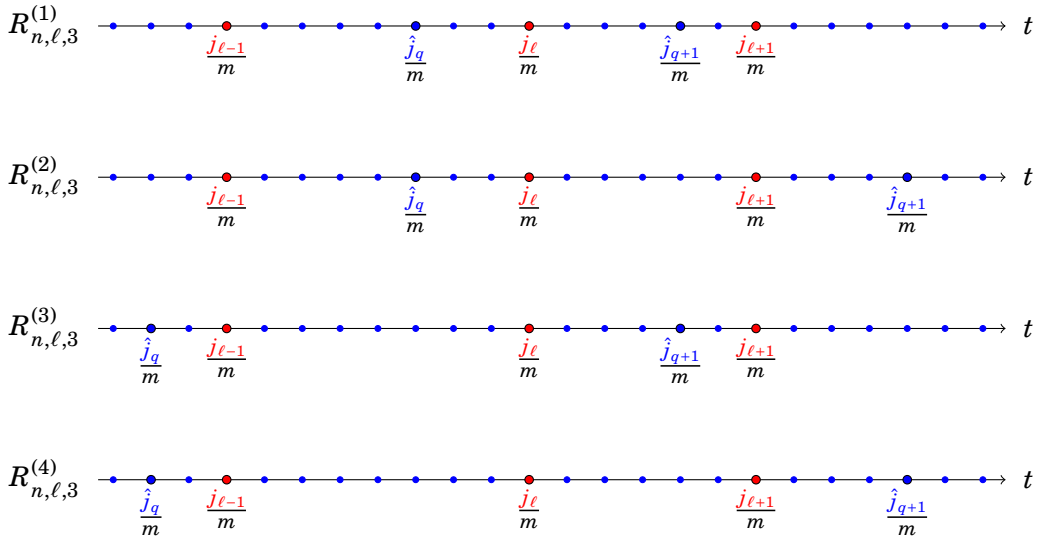


Fig. 2.11 – Illustration of the events $R^{(s)}_{n,\ell,3}$ for $s = 1,\ldots,4$.

with $j = j_\ell$ and $j = \hat{j}_q + 1$, respectively, which gives with probability one

$$\left| (j_\ell - \hat{j}_q - 2)(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}) + \sqrt{m}\bar{M}_n(\hat{j}_q + 1; j_\ell - 1) \right| \leq \hat{w}_{\hat{j}_q+1,j_\ell}. \tag{2.41}$$

Thus,

$$\left|\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}\right| \leq \frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{j_\ell - \hat{j}_q - 2} + \left|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q + 1; j_\ell - 1)}{j_\ell - \hat{j}_q - 2}\right|.$$

Second, we use (2.31) in Lemma 2.7.1 with $j = j_\ell + 1$ and $j = \hat{j}_{q+1}$, respectively, to get with probability one

$$\left| (\hat{j}_{q+1} - j_\ell - 2)\{(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m})\} + \sqrt{m}\bar{M}_n(j_\ell + 1; \hat{j}_{q+1} - 1) \right| \leq \hat{w}_{j_\ell+1,\hat{j}_{q+1}}.$$

Hence

$$\left|\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}\right| \leq \frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{\hat{j}_{q+1} - j_\ell - 2} + \left|\frac{\sqrt{m}\bar{M}_n(j_\ell + 1; \hat{j}_{q+1} - 1)}{\hat{j}_{q+1} - j_\ell - 2}\right|.$$

Define the event

$$
\begin{aligned}
Q^{(1)}_{n,\ell,3} = \Big\{ \big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big| \le \frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{|j_\ell - \hat{j}_q - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{j_\ell - \hat{j}_q}\Big| \\
+ \frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{|\hat{j}_{q+1} - j_\ell - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{\hat{j}_{q+1} - j_\ell - 2}\Big| \Big\} \\
\subset \Big\{ \big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_k-1,m}\big| \le \frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{m\varepsilon_n - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{m\varepsilon_n - 2}\Big| \\
+ \frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{m\varepsilon_n - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{m\varepsilon_n - 2}\Big| \Big\}.
\end{aligned}
$$

We observe that the event $Q^{(1)}_{n,\ell,3}$ occurs with probability one, so

$$
\begin{aligned}
\mathbb{P}[R^{(1)}_{n,\ell,3}] &= \mathbb{P}[R^{(1)}_{n,\ell,3} \bigcap Q^{(1)}_{n,\ell,3}] \\
&\le \mathbb{P}\Big[\frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{m\varepsilon_n - 2} \ge \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{m\varepsilon_n - 2} \ge \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{m\varepsilon_n - 2}\Big| \ge \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{m\varepsilon_n - 2}\Big| \ge \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big].
\end{aligned}
$$

Using Lemmas 2.7.1 and 2.7.2, and (2.16)-(2.17) from Assumption 2.4.3, each term of the last inequality goes to zero, as $n \to \infty$. For $\mathbb{P}[R^{(2)}_{n,\ell,3}]$, we apply Lemma 2.7.1 with $j = j_\ell$ and $j = \hat{j}_q + 1$ to obtain (2.41) and then with $j = j_\ell + 1$ and $j = j_{\ell+1}$ to get

$$
\Big|(j_{\ell+1} - j_\ell - 2)\{(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m})\} + \sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1}-1)\Big| \le \hat{w}_{j_\ell+1,j_{\ell+1}}.
$$

It follows that event $Q^{(2)}_{n,\ell,3}$ occurs with probability one, where

$$
\begin{aligned}
Q^{(2)}_{n,\ell,3} = \Big\{ \big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big| \le \frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{|j_\ell - \hat{j}_q - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{j_\ell - \hat{j}_q}\Big| \\
+ \frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{|j_{\ell+1} - j_\ell - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1}-1)}{j_{\ell+1} - j_\ell - 2}\Big| \Big\} \\
\subset \Big\{ \big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_k-1,m}\big| \le \frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{m\varepsilon_n - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{m\varepsilon_n - 2}\Big| \\
+ \frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{\Delta_{j,\min} - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{\Delta_{j,\min} - 2}\Big| \Big\}.
\end{aligned}
$$

Then

$$\begin{aligned}
\mathbb{P}[R^{(2)}_{n,\ell,3}] &= \mathbb{P}[R^{(2)}_{n,\ell,3} \bigcap Q^{(2)}_{n,\ell,3}] \\
&\leq \mathbb{P}\Big[\frac{\hat{w}_{\hat{j}_q+1,j_\ell}}{m\varepsilon_n - 2} \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{\Delta_{j,\min} - 2} \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(\hat{j}_q+1; j_\ell-1)}{m\varepsilon_n - 2}\Big| \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1}-1)}{\Delta_{j,\min} - 2}\Big| \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big].
\end{aligned}$$

Using Lemmas 2.7.1 and 2.7.2, (2.16)-(2.17) in Assumption 2.4.3, each term of the last inequality tends to zero as $n \to +\infty$. For $\mathbb{P}[R^{(3)}_{n,\ell,3}]$, we first use Lemma 2.7.1 with $j = j_{\ell-1}+1$ and $j = j_\ell$ to get

$$\Big|(j_\ell - j_{\ell-1} - 2)\big(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}\big) + \sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell - 1)\Big| \leq \hat{w}_{j_{\ell-1}+1,j_\ell}.$$

And then with $j = j_\ell + 1$ and $j = \hat{j}_{q+1}$, to obtain

$$\Big|(\hat{j}_{q+1} - j_\ell - 2)\big(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}\big) + \sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1} - 1)\Big| \leq \hat{w}_{j_\ell+1,\hat{j}_{q+1}}.$$

Hence the event $Q^{(3)}_{n,\ell,3}$ occurs with probability one, where

$$\begin{aligned}
Q^{(3)}_{n,\ell,3} &= \Big\{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big| \leq \frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{|j_\ell - j_{\ell-1} - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell-1)}{j_\ell - j_{\ell-1} - 2}\Big| \\
&\qquad + \frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{|\hat{j}_{q+1} - j_\ell - 2|} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{\hat{j}_{q+1} - j_\ell - 2}\Big|\Big\} \\
&\subset \Big\{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_k-1,m}\big| \leq \frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{\Delta_{j,\min} - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell-1)}{\Delta_{j,\min} - 2}\Big| \\
&\qquad + \frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{m\varepsilon_n - 2} + \Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{m\varepsilon_n - 2}\Big|\Big\}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{P}[R^{(3)}_{n,\ell,3}] &= \mathbb{P}[R^{(3)}_{n,\ell,3} \bigcap Q^{(3)}_{n,\ell,3}] \\
&\leq \mathbb{P}\Big[\frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{\Delta_{j,\min} - 2} \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\frac{\hat{w}_{j_\ell+1,\hat{j}_{q+1}}}{m\varepsilon_n - 2} \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell-1)}{\Delta_{j,\min} - 2}\Big| \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big] \\
&\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; \hat{j}_{q+1}-1)}{m\varepsilon_n - 2}\Big| \geq \frac{\big|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}\big|}{4}\Big].
\end{aligned}$$

By Lemmas 2.7.1 and 2.7.2, and (2.16)-(2.17) in Assumption 2.4.3, it implies that each term of the last inequality tends to zero as $n \to +\infty$.. Finally, for $\mathbb{P}[R^{(4)}_{n,\ell,3}]$, we first use Lemma 2.7.1 with $j = j_{\ell-1}+1$ and $j = j_\ell$ to obtain

$$\Big|(j_\ell - j_{\ell-1} - 2)\big\{\big(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}\big)\big\} + \sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell - 1)\Big| \leq \hat{w}_{j_{\ell-1}+1,j_\ell}.$$

Second, we use Lemma 2.7.1 with $j = j_\ell + 1$ and $j = j_{\ell+1}$ to obtain

$$\left|(j_{\ell+1} - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_{q+1}-1,m}) + \sqrt{m}\bar{M}_n(j_\ell + 1; j_{\ell+1} - 1)\right| \le \hat{w}_{j_\ell+1,j_{\ell+1}}.$$

It follows that the event $Q^{(4)}_{n,\ell,3}$ occurs with probability one, where

$$\begin{aligned} Q^{(4)}_{n,\ell,3} = \Big\{ |\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}| \le \frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{|j_\ell - j_{\ell-1} - 2|} + \left|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell - 1)}{j_\ell - j_{\ell-1} - 2}\right| \\ + \frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{|j_{\ell+1} - j_\ell - 2|} + \left|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1} - 1)}{j_{\ell+1} - j_\ell - 2}\right| \Big\} \\ \subset \Big\{ |\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_k-1,m}| \le \frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{\Delta_{j,\min} - 2} + \left|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell - 1)}{\Delta_{j,\min} - 2}\right| \\ + \frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{\Delta_{j,\min} - 2} + \left|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1} - 1)}{\Delta_{j,\min} - 2}\right| \Big\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}[R^{(4)}_{n,\ell,3}] &= \mathbb{P}[R^{(4)}_{n,\ell,3} \bigcap Q^{(4)}_{n,\ell,3}] \\ &\le \mathbb{P}\left[\frac{\hat{w}_{j_{\ell-1}+1,j_\ell}}{\Delta_{j,\min} - 2} \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\right] \\ &\quad + \mathbb{P}\left[\frac{\hat{w}_{j_\ell+1,j_{\ell+1}}}{\Delta_{j,\min} - 2} \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\right] \\ &\quad + \mathbb{P}\left[\left|\frac{\sqrt{m}\bar{M}_n(j_{\ell-1}+1; j_\ell - 1)}{\Delta_{j,\min} - 2}\right| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\right] \\ &\quad + \mathbb{P}\left[\left|\frac{\sqrt{m}\bar{M}_n(j_\ell+1; j_{\ell+1} - 1)}{\Delta_{j,\min} - 2}\right| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{4}\right] \\ &\to 0. \end{aligned}$$

as $n \to \infty$. This concludes the proof of Theorem 2.4.5. □

Appendix 2.A Technical Lemmas for the oracle inequalities

Here we prove Proposition 2.6.1 and Lemmas 2.6.2, 2.7.1 and 2.7.2

2.A.1 Proof of Proposition 2.6.1

Fix $j \in \{1, \dots, m\}$. We have

$$\begin{aligned} U_j &= \frac{1}{n}\sum_{i=1}^n \int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(s) dM_i(s), \\ V_j &= n\langle U_j \rangle = \frac{1}{n}\int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(s)\lambda_0(s)d(s), \\ \hat{V}_j &= n[U_j] = \frac{1}{n}\sum_{i=1}^n \int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(s) dN_i(s). \end{aligned}$$

Classical Bernstein deviation inequality applied to U_j, see Van De Geer (1995), yields that

$$\mathbb{P}\left[|U_j| \ge \sqrt{2\theta z} + \frac{z}{3n}, \frac{1}{n}\int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(s)\lambda_0(s)d(s) \le \theta\right] \le 2e^{-z}. \tag{2.A.1}$$

for all $\theta > 0$, and $z > 0$. It follows that

$$\mathbb{P}\Big[|U_j| \geq \sqrt{\frac{2\theta z}{n}} + \frac{z}{3n}, V_j \leq \theta\Big] \leq 2e^{-z}. \tag{2.A.2}$$

By choosing $\theta = c_0(z+1)/n$, this gives

$$\mathbb{P}\Big[|U_j| \geq \Big(\sqrt{2c_0} + \frac{1}{3}\Big)\frac{z+1}{n}, V_j \leq \frac{c_0(z+1)}{n}\Big] \leq 2e^{-z}. \tag{2.A.3}$$

For any $0 < \eta < \theta < \infty$, we have

$$\Big\{|U_j| \geq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n}\Big\} \bigcap \Big\{\eta < V_j \leq \theta\Big\} \subset \Big\{|U_j| \geq \sqrt{\frac{2\theta z}{n}} + \frac{z}{3n}\Big\} \bigcap \Big\{\eta < V_j \leq \theta\Big\}.$$

Together with (2.A.2), we obtain

$$\mathbb{P}\Big[|U_j| \geq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n}, \eta < V_j \leq \theta\Big] \leq 2e^{-z}. \tag{2.A.4}$$

Now we want to replace V_j by the observable $\hat{V}_j$ in the deviation (2.A.2). Define $\widetilde{U}_j$ by

$$\begin{aligned} \widetilde{U}_j &= \hat{V}_j - V_j \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_0^1 \mathbb{1}_{(\frac{j-1}{m},1]}(s) dM_i(s). \end{aligned}$$

Now writing again (2.A.4) and using the same argument as before, we arrive at

$$\mathbb{P}\Big[|\widetilde{U}_j| \geq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n}, \eta < V_j \leq \theta\Big] \leq 2e^{-z}. \tag{2.A.5}$$

But, if V_j satisfies

$$|\widetilde{U}_j| \leq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n},$$

then it satisfies

$$V_j \leq 2\hat{V}_j + 2\Big(\frac{\theta}{\eta} + \frac{1}{3}\Big)\frac{z}{n},$$

and

$$\hat{V}_j \leq 2V_j + 2\Big(\frac{1}{3} + 2\sqrt{\frac{\theta}{\eta}\Big(\frac{\theta}{\eta} + \frac{1}{3}\Big)} + 2\frac{\theta}{\eta}\Big)\frac{z}{n},$$

simply using the fact that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ for any $a, A, b > 0$. This proves that

$$\begin{aligned} &\Big\{|U_j| \leq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n}\Big\} \bigcap \Big\{\Big|\widetilde{U}_j\Big| \leq \sqrt{\frac{2\theta V_j z}{\eta n}} + \frac{z}{3n}\Big\} \\ &\qquad \subset \Big\{|U_j| \leq 2\sqrt{\frac{\theta z}{\eta n}\hat{V}_j} + \Big(\frac{1}{3} + 2\sqrt{\frac{\theta}{\eta}\Big(\frac{\theta}{\eta} + \frac{1}{3}\Big)} + 2\frac{\theta}{\eta}\Big)\frac{z}{n}\Big\}. \end{aligned} \tag{2.A.6}$$

So, using (2.A.4) and (2.A.5), we obtain

$$\mathbb{P}\Big[|U_j| \geq 2\sqrt{\frac{\theta z}{n}\hat{V}_j} + \Big(\frac{1}{3} + 2\sqrt{\frac{\theta}{\eta}\Big(\frac{\theta}{\eta} + \frac{1}{3}\Big)}\Big)\frac{z}{n}, \eta < V_j \leq \theta\Big] \leq 4e^{-z}.$$

The inequality is similar to (2.A.4), where we replaced V_j by the observable $\hat{V}_j$. It remains to remove the event $\{\eta < V_j \le \theta\}$ from this inequality. First, recall that (2.A.3) holds, so we can work on the event $\{V_j > c_0(z+1)/n\}$ from now on. We use a peeling argument. Define, for $q \ge 0$:

$$\theta_q = c_0 \frac{(z+1)}{n}(1+\varepsilon)^q,$$

and use the following decomposition into disjoint sets:

$$\{V_j > \theta_0\} = \bigcup_{q \ge 0} \{\theta_q < V_j \le \theta_{q+1}\}.$$

We have

$$\mathbb{P}\Big[|U_j| \ge c_{1,\varepsilon}\sqrt{\frac{z}{n}\hat{V}_j} + c_{2,\varepsilon}\frac{z}{n}, \theta_q < V_j \le \theta_{q+1}\Big] \le 4e^{-z}$$

where

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon} \quad \text{and} \quad c_{2,\varepsilon} = 2\sqrt{(1+\varepsilon)\big(\frac{4}{3}+\varepsilon\big)} + \frac{1}{3}.$$

Let

$$h_j = c_h \log\log\big(\frac{V_j}{\theta_0} \vee e\big).$$

On the event

$$\Big\{|\widetilde{U}_j| \le \sqrt{\frac{2(1+\varepsilon)V_j(z+h_j)}{n}} + \frac{(z+h_j)}{3n}\Big\}$$

we have

$$V_j \le 2\hat{V}_j + 2\big(\frac{4}{3}+\varepsilon\big)\frac{z}{n} + 2\frac{\frac{4}{3}+\varepsilon}{n}c_h \log\log\big(\frac{V_j}{\theta_0} \vee e\big),$$

which entails, assuming that $ec_0 > 2((1+\varepsilon)+\frac{1}{3})c_h$,

$$V_j \le \frac{ec_0(z+1)}{ec_0(z+1) - 2(\frac{4}{3}+\varepsilon)c_h}\Big(2\hat{V}_j + 2\big(\frac{4}{3}+\varepsilon\big)\frac{z}{n}\Big),$$

where we used the fact that $\log\log z \le z/e - 1$ for any $z \ge e$. This entails, together with (2.A.6), the following embedding:

$$\begin{aligned}\Big\{|U_j| \le \sqrt{\frac{2(1+\varepsilon)(z+h_j)V_j}{n}} + \frac{z+h_j}{3n}\Big\} \bigcap \Big\{|\widetilde{U}_j| \le \sqrt{\frac{2(1+\varepsilon)(z+h_j)V_j}{n}} + \frac{(z+h_j)}{3n}\Big\} \\ \subset \Big\{|U_j| \ge c_{1,\varepsilon}\sqrt{\frac{z+\hat{h}_{n,z,j}}{n}\hat{V}_j} + c_{2,\varepsilon}\frac{z+\hat{h}_{n,z,j}}{n}\Big\},\end{aligned}$$

where

$$\hat{h}_{n,z,j} = c_h \log\log\Big(\frac{2en\hat{V}_j + 2e(\frac{4}{3}+\varepsilon)z}{ec_0(z+1) - 2(\frac{4}{3}+\varepsilon)c_h} \vee e\Big).$$

Now, using the previous embeddings together with (2.A.4) and (2.A.5) we obtain

$$\begin{aligned}
&\mathbb{P}\Big[|U_j| \geq c_{1,\varepsilon}\sqrt{\frac{z+\hat{h}_{n,z,j}}{n}\hat{V}_j} + c_{2,\varepsilon}\frac{z+\hat{h}_{n,z,j}}{n}, V_j > \theta_0\Big] \\
&\quad \leq \sum_{q\geq 0}\mathbb{P}\Big[|U_j| \geq \sqrt{\frac{2(1+\varepsilon)V_j(z+h_j)}{n}} + \frac{z+h_j}{3n}, \theta_q < V_j \leq \theta_{q+1}\Big] \\
&\qquad + \sum_{q\geq 0}\mathbb{P}\Big[|\widetilde{U}_j| \geq \sqrt{\frac{2(1+\varepsilon)V_j(z+h_j)}{n}} + \frac{z+h_j}{3n}, \theta_q < V_j \leq \theta_{q+1}\Big] \\
&\quad \leq 4\Big(e^{-z} + \sum_{q\geq 1} e^{-\left(z+c_h \log\log(\frac{V_j}{\theta_0})\right)}\Big) \\
&\quad = 4\Big(1 + \big(\log(1+\varepsilon)\big)^{-c_h}\sum_{q\geq 1} q^{-c_h}\Big)e^{-z}.
\end{aligned}$$

Then with (2.A.3), it implies that

$$\mathbb{P}\Big[|U_j| \geq c_{1,\varepsilon}\sqrt{\frac{z+\hat{h}_{n,z,j}}{n}\hat{V}_j} + c_{3,\varepsilon}\frac{z+1+\hat{h}_{n,z,j}}{n}\Big] \leq \Big(6 + 4\big(\log(1+\varepsilon)\big)^{-c_h}\sum_{q\geq 1} q^{-c_h}\Big)e^{-z},$$

where $c_{3,\varepsilon} = \sqrt{2\max\big(c_0, 2(1+\varepsilon)(\frac{4}{3}+\varepsilon)\big)} + \frac{1}{3}$. □

2.A.2 Proof of Lemma 2.6.2

Using the fact that the functions $\{\lambda_{j,m} : j = 1,\dots,m\}$ form a basis of Λ_m, and under Assumption 2.2.1, one can give the explicit form of $\lambda_{0,m}$ as following

$$\lambda_{0,m} = m\sum_{j=1}^{m}\sum_{\ell=1}^{L_0}\beta_{0,\ell}|J_\ell \bigcap I_{j,m}|\mathbb{1}_{I_{j,m}} = \sum_{j=1}^{m}\beta_{0,j,m}\lambda_{j,m}, \tag{2.A.7}$$

here $|A|$ is the Lebesgue measure of the set A and $\beta_{0,j,m} = \sqrt{m}\sum_{\ell=1}^{L_0}\beta_{0,\ell}|J_\ell \bigcap I_{j,m}|$. We remark that the intervals J_ℓ do not share the same boundaries as the smaller intervals $I_{j,m}$. Setting the sequence $(\bar{\ell}_j)_{j=0,\dots,m}$ be the sequence defining by:

$$\bar{\ell}_0 = 1, \text{ and } \bar{\ell}_j = \max\{\ell = 1,\dots,L_0 : J_\ell \bigcap I_{j,m} \neq \emptyset\}, \text{ for } j = 1,\dots,m.$$

Using the sequence $(\bar{\ell}_j)_{j=0,\dots,m}$, one has the expression of the functions λ_0 and $\lambda_{0,m}$ as follows:

$$\lambda_0 = \sum_{j=1}^{m}\sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j}\beta_{0,\ell}\mathbb{1}_{J_\ell \cap I_{j,m}},$$

and

$$\lambda_{0,m} = m\sum_{j=1}^{m}\sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j}\beta_{0,\ell}|J_\ell \bigcap I_{j,m}|\mathbb{1}_{I_{j,m}} = \sum_{j=1}^{m}\alpha_{0,j,m}\mathbb{1}_{I_{j,m}} = \sum_{j=1}^{m}\alpha_{0,j,m}\sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j}\mathbb{1}_{J_\ell \cap I_{j,m}},$$

where $\alpha_{0,j,m} = m\sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \beta_{0,\ell}|J_\ell \bigcap I_{j,m}|$. From the fact that $\{\mathbb{1}_{J_\ell \cap I_{j,m}} : j = 1,\dots,m \text{ and } \ell = 1,\dots,L_0\}$ is an orthogonal basis of Λ_m (with respect to the $\mathbb{L}^2$-norm), we obtain

$$\begin{aligned}
\|\lambda_0 - \lambda_{0,m}\|^2 &= \Big\| \sum_{j=1}^{m} \sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \Big(\beta_{0,\ell} - m \sum_{\ell'=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \beta_{0,\ell'}|J_{\ell'} \bigcap I_{j,m}|\Big)\mathbb{1}_{J_\ell \cap I_{j,m}} \Big\|_2^2 \\
&= \sum_{j=1}^{m} \sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \Big(\beta_{0,\ell} - m \sum_{\ell'=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \beta_{0,\ell'}|J_{\ell'} \bigcap I_{j,m}|\Big)^2 |J_\ell \bigcap I_{j,m}| \\
&= \sum_{j=1}^{m} \mathbb{1}_{[\bar{\ell}_j - \bar{\ell}_{j-1} > 0]} \sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \Big(\beta_{0,\ell} - m \sum_{\ell'=\bar{\ell}_{j-1}}^{\bar{\ell}_j} \beta_{0,\ell'}|J_{\ell'} \bigcap I_{j,m}|\Big)^2 |J_\ell \bigcap I_{j,m}| \\
&\le \sum_{j=1}^{m} \mathbb{1}_{[\bar{\ell}_j - \bar{\ell}_{j-1} > 0]} \sum_{\ell=\bar{\ell}_{j-1}}^{\bar{\ell}_j} (\bar{\ell}_j - \bar{\ell}_{j-1} + 1) \max_{\ell,\ell'=k_{j-1},\dots,k_j} \Big(\beta_{0,\ell} - \beta_{0,\ell'}\Big)^2 |J_\ell \bigcap I_{j,m}| \\
&\le \sum_{j=1}^{m} \mathbb{1}_{[\bar{\ell}_j - \bar{\ell}_{j-1} > 0]} (\bar{\ell}_j - \bar{\ell}_{j-1} + 1) \max_{\ell,\ell' \in \{\bar{\ell}_{j-1},\dots,\bar{\ell}_j\}} \Big(\beta_{0,\ell} - \beta_{0,\ell'}\Big)^2 |I_{j,m}| \\
&\le \frac{2(L_0 - 1)\Delta_{\beta,\max}^2}{m}.
\end{aligned}$$

This proves Lemma 2.6.2. □

2.A.3 Proof of Lemma 2.7.1

To prove Lemma 2.7.1, we invoke subdifferential calculus, see Bertsekas (1999). We first write our objective functional as

$$\Phi(\mu) = \frac{1}{2}\sum_{j=1}^{m} (\mathbf{N}_j - (\mathbf{T}\mu)_j)^2 + \sum_{j=1}^{m} \hat{w}_j |\mu_{j,m}|.$$

So a necessary and sufficient condition for a vector $\hat{\mu}$ in $\mathbb{R}^m$ to minimize the function Φ is that the zero vector in $\mathbb{R}^m$ belongs to the sub-differential of $\Phi(\mu)$ at the point $\hat{\mu}$, that is, the following optimality condition holds:

$$\text{for all } j = 1,\dots,m \begin{cases} \Big(\mathbf{T}^\top\big(\mathbf{N} - \mathbf{T}\hat{\mu}\big)\Big)_j = \hat{w}_j \text{sign}(\hat{\mu}_{j,m}), \text{ if } \hat{\mu}_{j,m} \neq 0, \\ \Big|\Big(\mathbf{T}^\top\big(\mathbf{N} - \mathbf{T}\hat{\mu}\big)\Big)_j\Big| \le \hat{w}_j \text{sign}(\hat{\mu}_{j,m}), \text{ if } \hat{\mu}_{j,m} = 0. \end{cases}$$

Using that $(\mathbf{T}^\top \mathbf{N})_j = \sum_{q=j}^{m} \mathbf{N}_q$ and that $(\mathbf{T}^\top \hat{\beta})_j = \sum_{q=j}^{m} \hat{\beta}_{q,m}$, since $\mathbf{T}$ is a $m \times m$ lower triangular matrix having all its nonzero elements equal to one. Now, for $q = 1,\dots,m$, we observe that

$$\begin{aligned}
\mathbf{N}_q &= \sqrt{m} \int_{I_{q,m}} \lambda_0(t)dt + \sqrt{m}\bar{M}_n(I_{q,m}) \\
&= \sqrt{m} \int_{I_{q,m}} (\lambda_0(t) - \lambda_{0,m}(t))dt + \sqrt{m} \int_{I_{q,m}} \lambda_{0,m} dt + \sqrt{m}\bar{M}_n(I_{q,m}) \\
&= \sqrt{m} \int_{I_{q,m}} \Big(\sum_{\ell=1}^{L_0} \beta_{0,\ell}\mathbb{1}_{J_\ell} - \sqrt{m}\sum_{j=1}^{m} \beta_{0,j,m}\mathbb{1}_{I_{j,m}}\Big)dt \\
&\qquad + \sqrt{m}\sum_{j=1}^{m} \beta_{0,j,m} \int_{I_{q,m}} \mathbb{1}_{I_{j,m}}(t)dt + \sqrt{m}\bar{M}_n(I_{q,m}) \\
&= \beta_{0,q,m} + \sqrt{m}\bar{M}_n(I_{q,m}),
\end{aligned}$$

and we get the desired result.

□

2.A.4 Proof of Lemma 2.7.2

For the first statement, we have by definition,

$$\begin{aligned}\left|\bar{M}_n(a;b)\right| &= \Big|\sum_{q=a}^{b}\bar{M}_n(I_{q,m})\Big| \\ &= \Big|\frac{1}{n}\sum_{i=1}^{n}\sum_{q=a}^{b}\int_0^1 I_{q,m}(t)dM_i(t)\Big| \\ &= \Big|\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbf{1}_{(\frac{a-1}{m},\frac{b}{m}]}(t)dM_i(t)\Big|.\end{aligned}$$

Moreover, using Bernstein's inequality, it follows that, for any $z,\alpha>0$,

$$\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbf{1}_{(\frac{a-1}{m},\frac{b}{m}]}(t)dM_i(t)\Big|\geq z, \\ \langle\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbf{1}_{(\frac{a-1}{m},\frac{b}{m}]}(t)dM_i(t)\rangle\leq\alpha\Big]\leq 2\exp\Big\{-\frac{z^2}{2\alpha+\frac{2}{3}\rho z}\Big\},$$

where ρ is a upper bound of $\left\|\frac{1}{n}\mathbf{1}_{(\frac{a-1}{m},\frac{b}{m}]}\right\|_\infty$. Here we can choose $\rho=\frac{1}{n}$ and

$$\alpha = n^{-1}\int_{\frac{a-1}{m}}^{\frac{b}{m}}\lambda_0(t)dt = n^{-1}\mathbb{E}\Big[\bar{N}_n\Big((\frac{a-1}{m},\frac{b}{m}]\Big)\Big].$$

Hence, we obtain the first statement. For the second one, recall that for any $a=2,\dots,m$ we have

$$\hat{w}_a = c_1\sqrt{\frac{m(x+\log m+\hat{h}_{n,x,a})\hat{V}_a}{n}}+c_2\frac{\sqrt{m}(x+1+\log m+\hat{h}_{n,x,a})}{n}.$$

Since each term of $\hat{w}_a$ is positive and taking in account the dominant one, we have

$$\Big\{\hat{w}_a^2\geq\frac{m\log m}{n}\Big(\xi-\int_{\mathbb{1}_{(\frac{a-1}{m},1]}}\lambda_0(t)dt\Big)\Big\}\subset\Big\{\hat{V}_a\geq\xi-\int_{\mathbb{1}_{(\frac{a-1}{m},1]}}\lambda_0(t)dt\Big\},$$

for all $\xi>0$. By the Doob-Meyer decomposition theorem, we get

$$\Big\{\hat{w}_a^2\geq\frac{m\log m}{n}\Big(\xi-\int_{\mathbb{1}_{(\frac{a-1}{m},1]}}\lambda_0(t)dt\Big)\Big\}\subset\{\bar{M}_n(a;1)\geq\xi\},$$

Finally, by applying the first statement, see (2.33), to $\bar{M}_n(a;1)$, we concludes the proof of Lemma 2.7.2. □

Appendix 2.B Case II in the proof of Theorem 2.4.4

Here we prove the second case of the proof of Theorem 2.4.4 which is quite similar to the first one with a careful choice of the bounded terms in the approximate change-points sequence while applying the KKT optimality conditions. As $m\varepsilon_n\geq 6$ for all $n\geq 1$, it yields that the event $\{j_\ell+2<\hat{j}_\ell\}$ a.s.

2.B.1 Step II.1. Prove: $\mathbb{P}[A_{n,\ell} \bigcap C_n] \to 0$, as $n \to \infty$.

Applying (2.31) in Lemma 2.7.1 with $j = j_\ell + 1$ and $j = \hat{j}_\ell$, we get

$$-\hat{w}_{j_\ell} \le \sum_{q=j_\ell+1}^{m} \mathbf{N}_q - \sum_{q=j_\ell}^{m} \hat{\beta}_{q,m} \le \hat{w}_{j_\ell+1},$$

and

$$-\hat{w}_{\hat{j}_\ell} \le \sum_{q=\hat{j}_\ell}^{m} \mathbf{N}_q - \sum_{q=\hat{j}_\ell}^{m} \hat{\beta}_{q,m} \le \hat{w}_{\hat{j}_\ell}.$$

It follows that

$$\Big| \sum_{q=j_\ell+1}^{\hat{j}_\ell-1} \beta_{0,q,m} + \sqrt{m}\bar{M}_n(I_{q,m}) - \sum_{q=j_\ell+1}^{\hat{j}_\ell-1} \hat{\beta}_{q,m} \Big| \le \hat{w}_{j_\ell+1,\hat{j}_\ell}.$$

The property of the vector $\hat{\beta}$ in Lemma 2.7.1 yields that

$$\Big|(\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \hat{\beta}_{\hat{j}_\ell-1,m}) + \sqrt{m}\bar{M}_n(j_\ell+1;\hat{j}_\ell-1)\Big| \le \hat{w}_{j_\ell+1,\hat{j}_\ell}.$$

Therefore, on $C_n \bigcap \{\hat{j}_\ell > j_\ell + 2\}$ we have

$$\begin{aligned}\Big|&(\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_\ell-1,m}) \\ &\quad + \sqrt{m}\bar{M}_n(j_\ell+1;\hat{j}_\ell-1) \\ &\quad + (\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m})\Big| \le \hat{w}_{j_\ell+1,\hat{j}_\ell}.\end{aligned}$$

Define the event

$$\begin{aligned}C'_{n,\ell} = \Big\{\Big|&(\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_\ell-1,m}) \\ &+ \sqrt{m}\bar{M}_n(j_\ell+1;\hat{j}_\ell-1) \\ &+ (\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m})\Big| \le \hat{w}_{j_\ell+1,\hat{j}_\ell}\Big\}.\end{aligned}$$

It follows that $C'_{n,\ell}$ occurs with probability one. We observe that, $m\varepsilon_n \ge 6$ for all n, entails that $\frac{m\varepsilon_n}{2} - 2 \ge \frac{m\varepsilon_n}{6}$. Then

$$\{|\hat{j}_\ell - j_\ell| > \frac{m\varepsilon_n}{2}\} \subset \{|\hat{j}_\ell - j_\ell - 2| > \frac{m\varepsilon_n}{2} - 2\} \subset \{|\hat{j}_\ell - j_\ell - 2| \ge \frac{m\varepsilon_n}{6}\}.$$

Therefore,

$$\begin{aligned}&\mathbb{P}[A_{n,\ell} \bigcap C_n] \\ &= \mathbb{P}\Big[\Big\{\Big|(\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_\ell-1,m} - \hat{\beta}_{\hat{j}_\ell-1,m}) + \sqrt{m}\bar{M}_n(j_\ell+1;\hat{j}_\ell-1) \\ &\qquad + (\hat{j}_\ell - j_\ell - 2)(\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m})\Big| \le \hat{w}_{j_\ell+1,\hat{j}_\ell}\Big\} \\ &\qquad \bigcap A_{n,\ell} \bigcap C_n \bigcap \{\hat{j}_\ell > j_\ell + 2\}\Big] \\ &\le \mathbb{P}\Big[\Big\{\frac{\hat{w}_{j_\ell+1,\hat{j}_\ell}}{\frac{m\varepsilon_n}{6}} \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big\} \bigcap \Big\{\hat{j}_\ell > j_\ell + 2\Big\}\Big] \\ &\quad + \mathbb{P}\Big[\Big\{|\hat{\beta}_{\hat{j}_\ell-1,m} - \beta_{0,j_\ell-1,m}| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big\} \bigcap C_n\Big] \\ &\quad + \mathbb{P}\Big[\Big|\frac{\sqrt{m}\bar{M}_n(j_\ell;\hat{j}_\ell-1)}{\hat{j}_\ell - j_\ell - 2}\Big| \ge \frac{|\beta_{0,j_{\ell+1}-1,m} - \beta_{0,j_\ell-1,m}|}{3}\Big] \\ &:= \mathbb{P}[A'_{n,\ell,1}] + \mathbb{P}[A'_{n,\ell,2}] + \mathbb{P}[A'_{n,\ell,3}].\end{aligned}$$

We have

$$\begin{aligned}\mathbb{P}[A'_{n,\ell,1}] &\leq \mathbb{P}\Big[\hat{w}_{j_\ell+1,\hat{j}_\ell} \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18}\Big] \\ &\leq \mathbb{P}\Big[\hat{w}_{j_\ell+1} \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{36}\Big] \\ &= \mathbb{P}\Big[\hat{w}^2_{j_\ell+1} \geq \frac{m^2\varepsilon_n^2\Delta^2_{\beta,\min}}{36^2}\Big].\end{aligned}$$

By (2.16) in Assumption 2.4.3, and (2.34) in Lemma 2.7.2 with $\xi = \frac{nm\varepsilon_n^2\Delta^2_{\beta,\min}}{36^2\log m} + \mathbb{E}[\bar{N}_n((\frac{j_\ell}{m},1])]$, it follows that

$$\mathbb{P}(A'_{n,\ell,1}) \leq 2\exp\Big\{-\frac{n\xi^2}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell}{m},1]\Big)\Big]+\frac{2}{3}\xi}\Big\} \to 0,$$

as $n\to\infty$. Now,

$$A'_{n,\ell,3} \subset \Big\{\Big|\bar{M}_n(j_\ell;\hat{j}_\ell-1)\Big| \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18\sqrt{m}}\Big\} \subset \bigcup_{q=j_\ell+2}^{j_{\ell+1}-2}\Big\{\Big|\bar{M}_n(j_\ell;q)\Big| \geq \frac{m\varepsilon_n\Delta_{\beta,\min}}{18\sqrt{m}}\Big\}$$

Let $\varphi'_n := \frac{\sqrt{m}\varepsilon_n\Delta_{\beta,\min}}{18}$. Hence, by (2.33) in Lemma 2.7.2 we get

$$\begin{aligned}\mathbb{P}[A'_{n,\ell,3}] &\leq 2\sum_{q=j_\ell+2}^{j_{\ell+1}-2}\exp\Big\{-\frac{n\varphi'^2_n}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell-1}{m},\frac{q}{m}]\Big)\Big]+\frac{2}{3}\varphi'_n}\Big\} \\ &\leq 2(j_{\ell+1}-j_\ell-3)\exp\Big\{-\frac{n\varphi'^2_n}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell-1}{m},\frac{j_{\ell+1}-2}{m}]\Big)\Big]+\frac{2}{3}\varphi'_n}\Big\} \\ &\leq 2\exp\Big\{-\frac{n\varphi'^2_n}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_\ell-1}{m},\frac{j_{\ell+1}-2}{m}]\Big)\Big]+\frac{2}{3}\varphi'_n}+\log m\Big\}.\end{aligned}$$

By (2.16) in Assumption 2.4.3, we get $\mathbb{P}[A'_{n,\ell,3}]$ tends to zero as n tends to infinity. Let us now address $\mathbb{P}[A'_{n,\ell,2}]$. Using (2.31) in Lemma 2.7.1 with $j = j_\ell$ and with $j = \lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil$, and using the triangle inequality, it follows that

$$\Big|\sum_{q=\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil}^{j_\ell-1}\mathbf{N}_q - \sum_{q=\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil}^{j_\ell-1}\hat{\beta}_{q,m}\Big| \leq \hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell}.$$

On the event $C_n\cap\{\hat{j}_\ell > j_\ell\}$, we get

$$\Big|\frac{j_\ell-j_{\ell-1}}{2}(\beta_{0,j_\ell-1,m}-\hat{\beta}_{\hat{j}_\ell-1,m}) + \sqrt{m}\bar{M}_n(\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil;j_\ell-1)\Big| \leq \hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell}.$$

This implies

$$|\frac{j_\ell-j_{\ell-1}}{2}||\hat{\beta}_{\hat{j}_\ell-1,m}-\beta_{0,j_\ell-1,m}| \leq \hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell} + \Big|\sqrt{m}\bar{M}_n(\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil;j_\ell-1)\Big|.$$

Therefore, we may upper bound $\mathbb{P}[A'_{n,\ell,2}]$ as follows

$$
\begin{aligned}
&\mathbb{P}[A'_{n,\ell,2}] \\
&\quad= \mathbb{P}\Big[\Big\{|\hat{\beta}_{\hat{j}_\ell-1,m}-\beta_{0,j_\ell-1,m}| \geq \frac{|\beta_{0,j_{\ell+1}-1,m}-\beta_{0,j_\ell-1,m}|}{3}\Big\}\bigcap C_n\bigcap\{\hat{j}_\ell > j_\ell\}\Big] \\
&\quad= \mathbb{P}\Big[\Big\{|\frac{j_\ell-j_{\ell-1}}{2}||\hat{\beta}_{\hat{j}_\ell-1,m}-\beta_{0,j_\ell-1,m}| \\
&\qquad\qquad\qquad\qquad \geq |\frac{j_\ell-j_{\ell-1}}{2}|\frac{|\beta_{0,j_{\ell+1}-1,m}-\beta_{0,j_\ell-1,m}|}{3}\Big\}\bigcap C_n\Big] \\
&\quad\leq \mathbb{P}\Big[\Big\{\hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell}+\Big|\sqrt{m}\bar{M}_n(\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil;j_\ell-1)\Big| \\
&\qquad\qquad\qquad\qquad \geq |\frac{j_\ell-j_{\ell-1}}{2}|\frac{|\beta_{0,j_{\ell+1}-1,m}-\beta_{0,j_\ell-1,m}|}{3}\Big\}\bigcap C_n\Big] \\
&\quad\leq \mathbb{P}\Big[\hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell} \geq (j_\ell-j_{\ell-1})\frac{|\beta_{0,j_{\ell+1}-1,m}-\beta_{0,j_\ell-1,m}|}{6}\Big] \\
&\qquad+\mathbb{P}\Big[\Big|\sqrt{m}\bar{M}_n(\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil;j_\ell-1)\Big| \geq \frac{|\beta_{0,j_{\ell+1}-1,m}-\beta_{0,j_\ell-1,m}|}{6}\Big] \\
&\quad\leq \mathbb{P}\Big[\hat{w}_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil,j_\ell} \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{12}\Big]+\mathbb{P}\Big[\Big|\bar{M}_n(\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil;j_\ell-1)\Big| \geq \frac{\Delta_{j,\min}\Delta_{\beta,\min}}{12\sqrt{m}}\Big] \\
&\quad:= {\alpha'_{n,\ell,2}}^{(1)}+{\alpha'_{n,\ell,2}}^{(2)}.
\end{aligned}
$$

We observe that

$${\alpha'_{n,\ell,2}}^{(1)} \leq \mathbb{P}\Big[\hat{w}^2_{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil} \geq \frac{\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{24^2}\Big] \leq \mathbb{P}\Big[\hat{w}^2_{j_{\ell-1}} \geq \frac{\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{24^2}\Big].$$

By (2.17) in Assumption 2.4.3, (2.34) in Lemma 2.7.2 with $\xi = \frac{n\Delta^2_{j,\min}\Delta^2_{\beta,\min}}{24^2 m\log m}+\mathbb{E}[\bar{N}_n((\frac{j_{\ell-1}-1}{m},1])]$, it follows that

$${\alpha'_{n,\ell,2}}^{(1)} \quad\leq\quad 2\exp\Big\{-\frac{n\xi^2}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{j_{\ell-1}-1}{m},1]\Big)\Big]+\frac{2}{3}\xi}\Big\}\to 0,$$

as $n\to\infty$. By (2.33) in Lemma 2.7.2 with $z=\frac{\Delta_{j,\min}\Delta_{\beta,\min}}{12\sqrt{m}}$ and (2.17) in Assumption 2.4.3, we obtain

$${\alpha'_{n,\ell,2}}^{(2)} \quad\leq\quad 2\exp\Big\{-\frac{nz^2}{2\mathbb{E}\Big[\bar{N}_n\Big((\frac{\lceil\frac{j_\ell+j_{\ell-1}}{2}\rceil-1}{m},\frac{j_\ell-1}{m}]\Big)\Big]+\frac{2}{3}z}\Big\}\to 0,$$

as $n\to\infty$. Therefore, we conclude that $\mathbb{P}[A'_{n,\ell,2}]\to 0, as\, n\to\infty$.

2.B.2 Step II.2. Prove: $\mathbb{P}[A_{n,\ell}\bigcap C_n^{\complement}]\to 0, as\, n\to\infty$.

As in Case I from Section 2.7.1, we split $\mathbb{P}[A_{n,\ell}\bigcap C_n^{\complement}]$ into

$$\mathbb{P}[A_{n,\ell}\bigcap C_n^{\complement}] = \mathbb{P}[A_{n,\ell}\bigcap D_n^{(l)}]+\mathbb{P}[A_{n,\ell}\bigcap D_n^{(m)}]+\mathbb{P}[A_{n,\ell}\bigcap D_n^{(r)}].$$

Let us first focus on $\mathbb{P}[A_{n,\ell}\bigcap D_n^{(m)}]$, see Figure 2.B.1. Note that

$$
\begin{aligned}
\mathbb{P}[A_{n,\ell}\bigcap D_n^{(m)}\bigcap\{\hat{j}_\ell > j_\ell\}] &\leq \mathbb{P}[A_{n,\ell}\bigcap B_{\ell+1,\ell}\bigcap D_n^{(m)}] \\
&\quad+\sum_{l=\ell+1}^{L_0-2}\mathbb{P}[C_{l,l}\bigcap B_{l+1,l}\bigcap D_n^{(m)}].
\end{aligned}
\tag{2.B.1}
$$

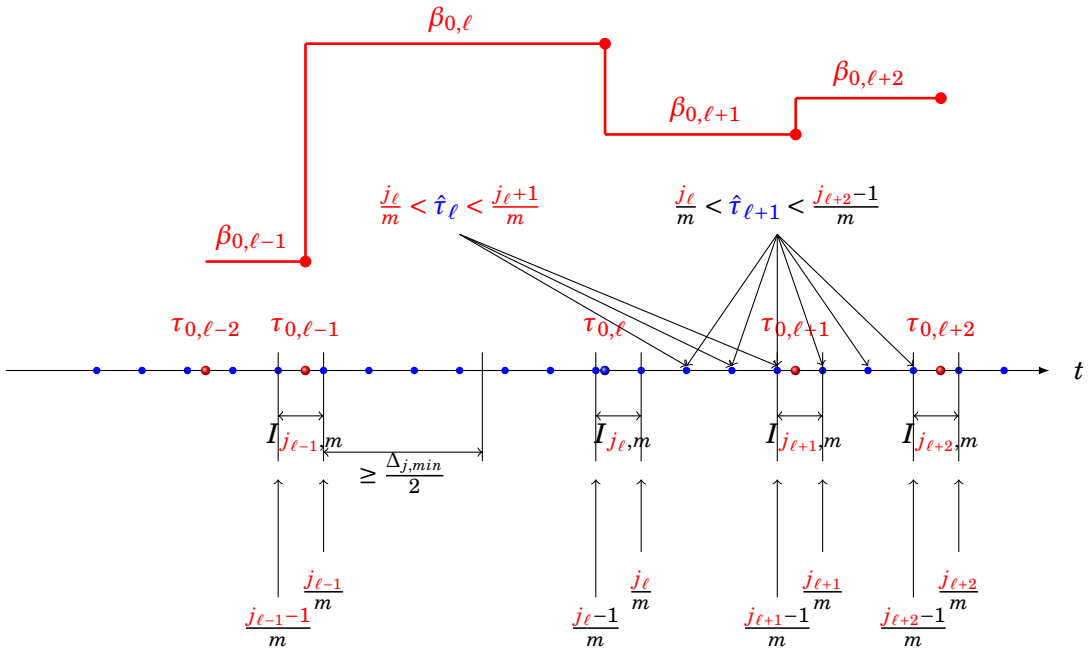


Fig. 2.B.1 – A zoom into the Case II.

Let us now prove that the first term in the right hand side of (2.B.1) goes to zero as n tends to infinity. Using (2.31) in Lemma 2.7.1 with $j = \lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil$ and $j = j_{\ell+1}$, on the first hand and (2.31) in Lemma 2.7.1 with $j = j_{\ell+1}+1$ and with $j = j_{\ell+2}$ on the other hand, we obtain, respectively

$$|\frac{j_{\ell+1}-j_\ell}{2}||\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \leq \hat{w}_{\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil, j_{\ell+1}} + |\sqrt{m}\bar{M}_n(\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil; j_{\ell+1}-1)|, \tag{2.B.2}$$

and

$$|j_{\ell+2}-j_{\ell+1}-2||\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+2}-1,m}| \leq \hat{w}_{j_{\ell+1}+1,j_{\ell+2}} + |\sqrt{m}\bar{M}_n(j_{\ell+1}+1; j_{\ell+2}-1)|. \tag{2.B.3}$$

In addition, we have

$$\begin{aligned}
&|\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \\
&\quad = |(\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}) - (\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+2}-1,m})| \\
&\quad \leq |\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+1}-1,m}| + |\hat{\beta}_{\hat{j}_{\ell+1}-1,m} - \beta_{0,j_{\ell+2}-1,m}| \\
&\quad \leq \frac{\hat{w}_{\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil, j_{\ell+1}}}{|\frac{j_{\ell+1}-j_\ell}{2}|} + \frac{\sqrt{m}\bar{M}_n(\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil; j_{\ell+1}-1)|}{|\frac{j_{\ell+1}-j_\ell}{2}|} \\
&\qquad + \frac{\hat{w}_{j_{\ell+1}+1,j_{\ell+2}}}{|j_{\ell+2}-j_{\ell+1}-2|} + \frac{|\sqrt{m}\bar{M}_n(j_{\ell+1}+1; j_{\ell+2}-1)|}{|j_{\ell+2}-j_{\ell+1}-2|} \\
&\quad \leq 2\frac{\hat{w}_{\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil, j_{\ell+1}}}{\Delta_{j,\min}} + 2\frac{\sqrt{m}\bar{M}_n(\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil; j_{\ell+1}-1)|}{\Delta_{j,\min}} \\
&\qquad + \frac{\hat{w}_{j_{\ell+1}+1,j_{\ell+2}}}{|\Delta_{j,\min}-2|} + \frac{|\sqrt{m}\bar{M}_n(j_{\ell+1}+1; j_{\ell+2}-1)|}{|\Delta_{j,\min}-2|}.
\end{aligned}$$

Define the event $E'_{n,\ell}$ by

$$E'_{n,\ell} = \Big\{ |\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}| \leq \frac{2\hat{w}_{\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil, j_{\ell+1}}}{\Delta_{j,\min}} + \frac{6\hat{w}_{j_{\ell+1}+1,j_{\ell+2}}}{\Delta_{j,\min}} + \frac{2\sqrt{m}\bar{M}_n(\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil; j_{\ell+1}-1)|}{\Delta_{j,\min}} + \frac{|6\sqrt{m}\bar{M}_n(j_{\ell+1}+1; j_{\ell+2}-1)|}{\Delta_{j,\min}} \Big\}.$$

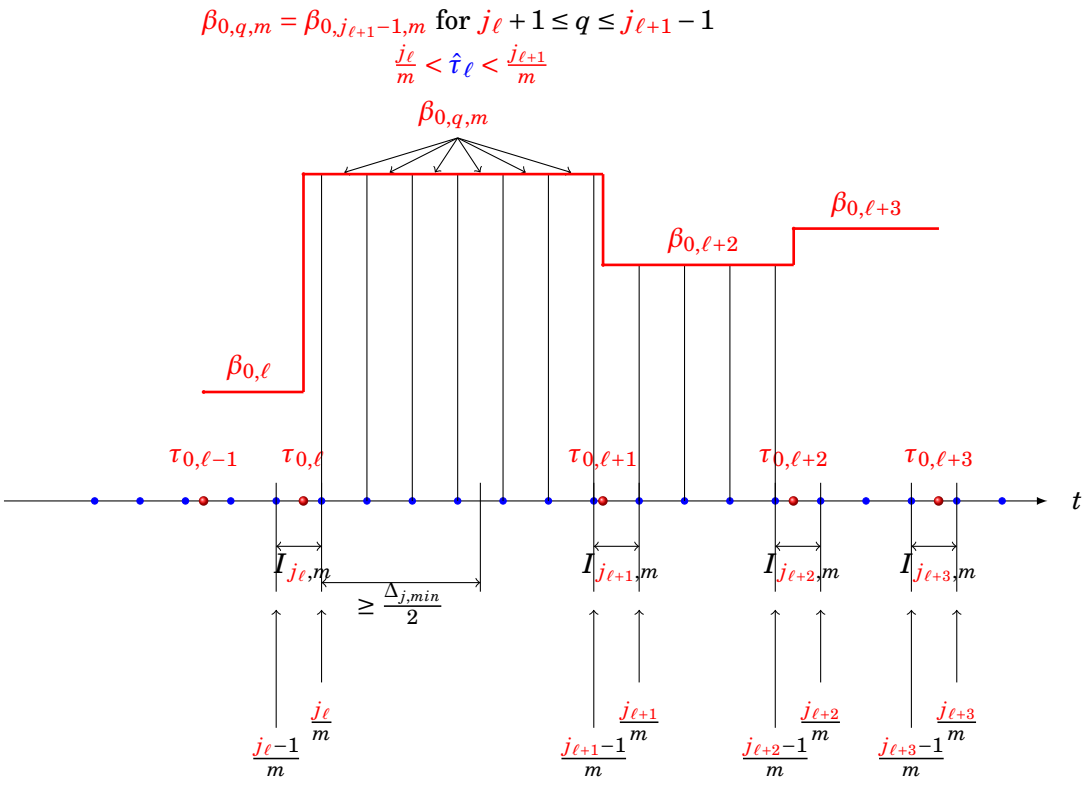


Fig. 2.B.2 – A zoom of $\beta_{0,q,m}$, the coefficients of the projection function $\lambda_{0,m}$ in Case II.

$E'_{n,\ell}$ occurs with probability one, see Figure 2.B.2. Therefore, we obtain

$$\begin{aligned}
&\mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)}] \\
&\quad \leq \mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)} \bigcap E'_{n,\ell}] \\
&\quad \leq \mathbb{P}\Big[\hat{w}_{\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil, j_{\ell+1}} \geq \frac{\Delta_{j,\min}|\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}|}{8}\Big] \\
&\qquad + \mathbb{P}\Big[\hat{w}_{j_{\ell+1}+1,j_{\ell+2}} \geq \frac{|\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}|}{24}\Big] \\
&\qquad + \mathbb{P}\Big[|\bar{M}_n(\lceil \frac{j_{\ell+1}+j_\ell}{2} \rceil; j_{\ell+1}-1)| \geq \Delta_{j,\min}\frac{|\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}|}{8\sqrt{m}}\Big] \\
&\qquad + \mathbb{P}\Big[|\bar{M}_n(j_{\ell+1}+1; j_{\ell+2}-1)| \geq \Delta_{j,\min}\frac{|\beta_{0,j_{\ell+2}-1,m} - \beta_{0,j_{\ell+1}-1,m}|}{24\sqrt{m}}\Big] \\
&\quad := \theta'_{n,\ell,1} + \theta'_{n,\ell,2} + \theta'_{n,\ell,3} + \theta'_{n,\ell,4}.
\end{aligned}$$

By (2.33)-(2.34) in Lemma 2.7.2, and (2.16)-(2.17) in Assumption 2.4.3, we show that for $s = 1,\dots,4, \theta'_{n,\ell,s} \to 0$, $\mathbb{P}[A_{n,\ell} \bigcap B_{\ell+1,\ell} \bigcap D_n^{(m)}] \to 0$, as $n \to \infty$. Recall that in Case I from Section 2.7.1, we proved $\mathbb{P}[A_{n,\ell} \bigcap D_n^{(l)}] \to 0$, as $n \to \infty$ and in a similar way $\mathbb{P}[A_{n,\ell} \bigcap D_n^{(r)}] \to 0$, as $n \to \infty$. This concludes the proof of Theorem 2.4.4. □

Chapter 3

Binarsity: Features Binarization and Cuts Selection using Convex Optimization

This chapter is a preprint of Alaya et al. (2016b).

Abstract

In the present paper, we deal with the problem of estimation regression function for generalized linear models in high-dimensional settings. Towards this end, we introduce a new notion of sparsity called *binarsity*. It computes the different values of the parameter extended in a space of binarized features. We focus on an estimation procedure based on a weighted data-driven of binarsity. We get convergence rates for the prediction error by proving non-asymptotic oracle inequalities promoting binarsity sparsity of the covariables. We give an algorithm that efficiently solves the convex problem studied in this work.

Contents

3.1 Introduction

The challenges of high-dimensionality arise in diverse fields of sciences, ranging from computational biology and health studies to financial engineering and risk management, etc.. The availability of massive data along with new scientific problems have presented serious challenges to existing learning methods and reshaped statistical thinking and data analysis. To address the problem of the curse of dimensionality, sparse inference is now an ubiquitous technique for dimension reduction and variable selection, see for instance Bühlmann and Van De Geer (2011), Hastie et al. (2001) among many others. The principle of the sparsity is appropriate for model building and usually results in the most stable and robust model. A fundamental step in sparsity is to do careful variable selection, where the goal again is to build the best performing model for future and ongoing use. The idea of sparsity is to add a penalty term on model complexity to some model fitting loss measure. Then minimizing the penalized model fitting loss measure yields an estimate of the model parameters. The standard sparsity, namely a parameter that contains (unpatterned) zeros if typically enforced by the use of an ℓ_1-penalization, often called Lasso (see Tibshirani (1996)). It is a big breakthrough in the field of sparse model estimation which performs the variable selection and coefficient shrinkage simultaneously. It has been demonstrated in Tibshirani (1996) that the Lasso is more stable and accurate than traditional variable selection methods such as best subset selection. The statistical properties of the Lasso have been extensively investigated (see Bickel et al. (2009); Bunea et al. (2007); Knight and Fu (2000); Zhao and Yu (2006)).

One drawback of the Lasso is the fact that it ignores ordering of the features. Tibshirani et al. (2005) proposed the well known structured sparse model, fused Lasso, which provides superior performance in recovering the true model. Due to the fact that fused Lasso is a sum of the ℓ_1-norm and the total-variation penalty, it enforces sparsity in both the coefficients and their successive differences, which is desirable for applications with features ordered in some meaningful way. Fused Lasso has achieved great success in many applications such as comparative genomic hybridization (see Rapaport et al. (2008)), image denoising (see Friedman et al. (2007)), and prostate cancer analysis (see Tibshirani et al. (2005)), where features in the true model are closely related to their neighbors. Furthermore, the total-variation model was first introduced in Rudin et al. (1992) as a regularization approach to remove noise and handle proper edges in a given image. It is widely used in sparse signal (see Little and Jones (2011)), especially when it is known the signal to be recovered is piecewise constant (see Alaya et al. (2015); Harchaoui and Lévy-Leduc (2010)).

An important principle in machine learning is that one should not expect the model to do all the difficult work. In practical situations with noisy data, it is often useful to encode the inputs as best as possible using expert knowledge and good statistical practices (see Wu and Coggeshall (2012)). Yet, this generally reduces potential information of the model, but in practice it allows the model to focus its efforts in the right areas. One of the basic encoding techniques is *feature discretization*. Feature discretization partitions the range of a continuous feature into finite set of intervals and relates these intervals with meaningful labels. It can also be defined as a process

used to quantify continuous features (see Liu et al. (2002)). An advantage of feature discretization is that it can increase the accuracy of the prediction learning, the processing speed and produce results that are more compact, concise and accurate over continuous data (see Chapelle et al. (2014); Dougherty et al. (1995)).

In supervised learning, the discretization of a continuous features specifies the set of *cut-points* in the continuous range of the feature that delimit the intervals to be mapped into the discretized feature. To formalize the problem, a discretization algorithm would discretize a continuous feature in a data set, into r discrete intervals $\mathscr{P} = \{[t_0, t_1), [t_1, t_2), \ldots, [t_{r-1}, t_r]\}$, where t_0 is the minimal value, t_r is the maximal value and $t_l < t_{l+1}$, for $l = 0, 1, \ldots, r-1$. Such a partition $\mathscr{P}$ is called a discretization scheme on a feature and $\mathscr{T} = \{t_0, t_1, \ldots, t_r\}$ is the set of cut-points of feature, see Garcia et al..

A vast number of discretization techniques can be found in the literature, see Garcia et al.; Liu et al. (2002) for two recent overviews, and references therein for a survey of discretization techniques. Obtaining the optimal discretization is NP-hard problem (see Chlebus and Nguyen (1998)). Furthermore, the best discretization algorithm is the one which significantly reduces the number of discrete intervals of a continuous feature and maximizes the accuracy and efficiency of the learning tasks.

Find cuts for quantitative features can be generally achieved using classification algorithms based on tree: CART Breiman et al. (1984) applies the Gini criterion, i.e., a measure of the impurity of their intervals. C4.5 Quinlan (1993) uses the information gain based on Shannon entropy. However such an approach finds cuts for a features by looking at the influence on the output labels separately for each feature.

The contribution of this paper is to combine the features binarization technique, together with a total-variation regularizer on each binarized features, to find the relevant cuts in the extended parameters space of binarized features. This paper introduces a new notion of sparsity called *binarsity*, that penalizes the number of different values of the parameter extended in the space of binarized features.

The remainder of this paper is organized as follows. We begin in Section 3.2 with the set-up of the feature binarization technique which is the principal key tool to construct the binarsity penalty. Then, we consider a weighted version of binarsity. The weights are chosen with respect to the matrix design. We finish this section by presenting an algorithm that calculate the proximal operator of binarsity, see Algorithm 4. Section 3.3 is devoted to the learning procedures under binarsity scenario by introducing the appropriate convex problem to be studied. We establish oracles inequalities for the prediction error in the simple and generalized linear models, see Theorems 3.3.3 and 3.3.7. Section **??** contains simulations results. The proofs of all the results stated in the paper are deferred to Section 3.4.

Notation. Throughout the paper, for every $q \in [0, \infty]$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, that is

$$\|v\|_q = \begin{cases} \#\big(\{k : v_k \neq 0\}\big), & \text{if } q = 0, \\ \big(\sum_{k=1}^m |v_k|^q\big)^{1/q}, & \text{if } 0 < q < \infty, \\ \max_{k=1,\ldots,m} |v_k|, & \text{if } q = \infty. \end{cases}$$

For two vectors u and v of the same dimension m, we denote by $u \odot v$ the Hadamard product between u and v, that is $u \odot v = (u_1 v_1, \ldots, u_m v_m)^\top$. For any vector $u \in \mathbb{R}^m$, and any subset L of $\{1,\ldots,m\}$, u_L is the vector in $\mathbb{R}^m$ that has the same coordinates as u on L and zero coordinates on the complement $L^c = \{1,\ldots,m\} \backslash L$. We write $\mathbf{1}_m$ (resp. $\mathbf{0}_m$) for the vector of $\mathbb{R}^m$having all coordinates equal to one (resp. zero). Given two sets $\mathscr{E} = \{e_1,\ldots,e_r\}$ and $\mathscr{F} = \{f_1,\ldots,f_s\}$, we denote by $[\mathscr{E},\mathscr{F}]$ the concatenation of $\mathscr{E}$ and $\mathscr{F}$, namely $[\mathscr{E},\mathscr{F}] = \{e_1,\ldots,e_r,f_1,\ldots,f_s\}$. We write $|A|$ for the cardinality of the set A. Finally, we denote by $\mathrm{sign}(x)$ the sub-differential of the function $x \mapsto |x|$, that is

$$\mathrm{sign}(x) = \begin{cases} \{1\}, & \text{if } x > 0, \\ [-1,1], & \text{if } x = 0, \\ \{-1\}, & \text{if } x < 0. \end{cases}$$

3.2 Binarsity, cuts and convex optimization

In this section we give the set-up of the feature binarization technique which is the principal key tool to construct the binarsity penalty.

3.2.1 Features binarization

We have a raw design matrix $\boldsymbol{X} = [\boldsymbol{X}_{i,j}]_{1\le i\le n; 1\le j\le p}$ with n examples and p raw features. We denote by $\boldsymbol{X}_{\bullet,j}$ the j-th feature column and by $\boldsymbol{X}_{i,\bullet}$ the i-th data row of $\boldsymbol{X}$. The binarized matrix $\boldsymbol{X}^B$ is a matrix with an extended number d of columns, typically with d much larger than p, where the j-th column $\boldsymbol{X}_{\bullet,j}$ is replaced by a number d_j of columns $\boldsymbol{X}^B_{\bullet,j,1},\ldots,\boldsymbol{X}^B_{\bullet,j,d_j}$ containing only zeros and ones, see Figure 3.1. For a given raw feature j, we consider two cases: either we decide to treat the feature as qualitative or quantitative. If we choose feature j to be qualitative, assuming that the entries of $X_{\bullet,j}$ takes values (modalities) in the set $\{1,\ldots,M_j\}$ with cardinality M_j, we take $d_j = M_j$, and use a binary coding of each modality by defining

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1, & \text{if } \boldsymbol{X}_{i,j} = k, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1,\ldots,n$ and $k = 1,\ldots,d_j$. If we choose feature j to be quantitative, then we quantize it to a smaller number of values d_j. We consider a partition of intervals $I_{j,1},\ldots,I_{j,d_j}$ for the range of values of $\boldsymbol{X}_{\bullet,j}$ and define

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1, & \text{if } \boldsymbol{X}_{i,j} \in I_{j,k}, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1,\ldots,n$ and $k = 1,\ldots,d_j$. A natural choice of intervals is given by the quantiles, namely we can typically choose $I_{j,k} = \left[q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right)$ for $k = 1,\ldots,d_j - 1$, $I_{j,d_j} = \left[q_j(\frac{d_j-1}{d_j}), q_j(1)\right]$, and where $q_j(\alpha)$ denotes a quantile of order α for $\boldsymbol{X}_{\bullet,j}$.

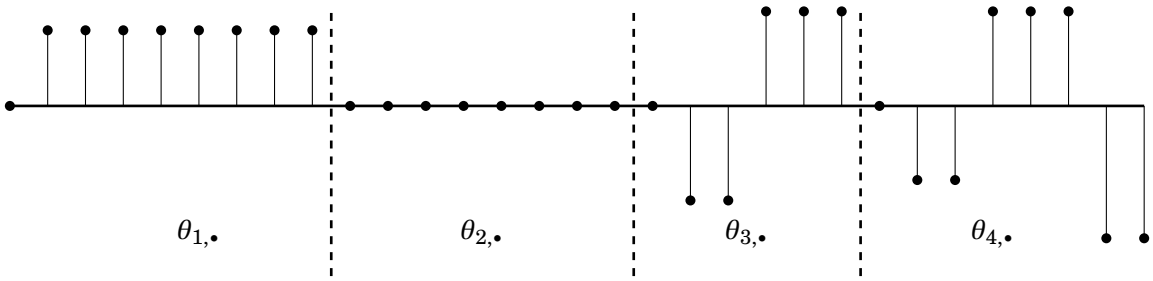


Fig. 3.1 – Illustration of $\theta = (\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top)^\top$ with: $p = 4, d_1 = 9$, $d_2 = 8$, $d_3 = 6$, $d_4 = 8$.

3.2.2 Binarsity

To each binarized feature $\boldsymbol{X}^B_{\bullet,j,k}$ corresponds a parameter $\theta_{j,k}$. The parameters associated to the binarization of the j-th feature is denoted $\theta_{j,\bullet} = (\theta_{j,1} \cdots \theta_{j,d_j})^\top$. The full parameters vector of size $d = \sum_{j=1}^p d_j$, is simply obtained by concatenation of the blocks $\theta_{j,\bullet}$ of each feature j, namely

$$\theta = (\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top)^\top = \big(\theta_{1,1} \cdots \theta_{1,d_1} \theta_{2,1} \cdots \theta_{2,d_2} \cdots \theta_{p,1} \cdots \theta_{p,d_p}\big)^\top.$$

Important remark. The binarized matrix $\boldsymbol{X}^B$ has less than full rank, since in each block the sum of the columns $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j}$ is equal to $\mathbf{1}_n$ (intercept). Hence, the parameter corresponds to the intercept is intrinsically aliased to $\sum_{k=1}^{d_j} \theta_{j,k}$. To avoid this over-parametrization, we must add a constraint. There is no unique way to do this, and the choice generally depends on interpretability. We can either drop a parameter or add a linear constraint in each bloc $\theta_{j,\bullet}$. That is, one sets $\theta_{j,k} = 0$, for one value k in $\{1, \ldots, d_j\}$. This is called a k^{th}-baseline- (or reference) constraint. Another useful possibility is to impose $\sum_{k=1}^{d_j} \theta_{j,k} = 0$, called sum-to-zero-constraint (see Agresti (2015)). In our case, we want each block $\theta_{j,\bullet}$ to be either constant, or to contain a small number of different values. Therefore, our choice consists in working under the sum-to-zero-constraint. This allows to detect cuts, and to have a different parameters for different values of the raw feature.

The notion of sparsity we therefore introduce, called *binarsity* is computed as

$$\begin{aligned} \text{bina}(\theta) &= \sum_{j=1}^p \Big(\|\theta_{j,\bullet}\|_{\text{TV}} + \delta_{\mathcal{H}_j}(\theta_{j,\bullet}) \Big) \\ &= \sum_{j=1}^p \Big(\sum_{k=2}^{d_j} |\theta_{j,k} - \theta_{j,k-1}| + \delta_{\mathcal{H}_j}(\theta_{j,\bullet}) \Big), \end{aligned}$$

where $\|\cdot\|_{\text{TV}}$ is the total variation penalization, $\mathcal{H}_j = \{\beta_{j,\bullet} \in \mathbb{R}^{d_j} : \sum_{k=1}^{d_j} \beta_{j,k} = 0\}$ (the hyperplan of $\mathbb{R}^{d_j}$ with normal vector $\mathbb{1}_{j,\bullet} = (1, \ldots, 1)^\top$), and the indicator function

$$\delta_{\mathcal{H}_j}(\beta_{j,\bullet}) = \begin{cases} 0, & \text{if } \beta_{j,\bullet} \in \mathcal{H}_j, \\ \infty, & \text{otherwise.} \end{cases}$$

If a raw feature j is statistically not relevant for predicting the labels, then the full block $\theta_{j,\bullet}$ should be constant, and in this case the contribution of this block to binarsity is zero. If the raw feature j is relevant, then the number of different values for the coefficients of $\theta_{j,\bullet}$ should be kept as small as possible, in order to balance bias and variance.

The intuition behind the introduction of this penalty was the desire to produce a piecewise constant coefficient in the penalized estimator. Based on the analysis given below, we consider the following data-driven weighted version of this penalty given by

$$\begin{aligned}\mathrm{bina}_{\hat{w}}(\theta) &= \sum_{j=1}^{p} \Big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}(\theta_{j,\bullet}) \Big) \\ &= \sum_{j=1}^{p} \Big(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| + \delta_{\mathscr{H}_j}(\theta_{j,\bullet}) \Big)\end{aligned}$$

where $\hat{w} = (\hat{w}_{1,\bullet} \dots \hat{w}_{p,\bullet})$ is a data-driven block vector of weights, chosen with respect to the design matrix $\boldsymbol{X}$, such that for all $j = 1, \dots, p$, $\hat{w}_{j,1} = 0$ and for all $k \in \{2, \dots, d_j\}$

$$\hat{w}_{j,k} \approx \sqrt{\frac{d_{\max} \hat{n}_{j,k}}{n}},$$

where $d_{\max} = \max_{j=1,\dots,p} d_j$, and

$$\hat{n}_{j,k} = \frac{\#\Big(\Big\{ i = 1, \dots, n : \boldsymbol{X}_{i,j} \in \big[q_j\big(\tfrac{k}{d_j}\big), q_j(1) \big] \Big\} \Big)}{n}.$$

We observe that $\hat{n}_{j,k}$ is the number of 1 in the sub-matrix obtained by deleting the $(k-1)$ columns in the j-th binarized block matrix $\boldsymbol{X}^B_{\bullet,j}$.

3.2.3 Proximal operator of binarsity

Knowing the proximal operator for a penalization function immediately provides possibilities for efficient algorithms for solving the penalized regularization problem. To begin, we recall here the definition of the proximal operator, see Bauschke and Combettes (2011). Let φ be a real valued convex function on $\mathbb{R}^d$. The proximal operator of φ is defined, for every $v \in \mathbb{R}^d$ by $\mathrm{prox}_{\varphi}(v) := \mathrm{argmin}_{u \in \mathbb{R}^d} \left\{ \frac{1}{2} \|v - u\|_2^2 + \varphi(u) \right\}$. Proximal operators can be interpreted as generalized projections because if φ is the indicator function of a convex set, then $\mathrm{prox}_{\varphi}(v)$ is the projection of v onto the set. Generally the proximal operator is easy to calculate when the function φ is separable, see Bach et al. (2012). Next, we proceed to compute the proximal operator of weighted binarsity. Since this penalty is separable by blocks, we have

$$\big(\mathrm{prox}_{\mathrm{bina}_{\hat{w}}}(\theta) \big)_{j,\bullet} = \mathrm{prox}_{(\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j})}(\theta_{j,\bullet}),$$

for all $j = 1, \dots, p$. Thus, let us focus on a single j-th block. Proposition 3.2.1 below expresses $\mathrm{prox}_{\mathrm{bina}_{\hat{w}}}$ based on the proximal operator of the weighted total-variation penalization, namely $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}$. We refer to Alaya et al. (2015) where the authors gave an algorithm for $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}$.

Algorithm 4:

for $j = 1, \dots, p$ **do**
  $\beta_{j,\bullet} \leftarrow \mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})$;
  $\eta_{j,\bullet} \leftarrow \beta_{j,\bullet} - \frac{1}{d_j}\sum_{k=1}^{d_j}\beta_{j,k}$;
  return $\eta_{j,\bullet}$;

Proposition 3.2.1. *Algorithm 4 gives* $\mathrm{prox}_{\mathrm{bina}_{\hat{w}}}$.

The proof of Proposition 3.2.1 is carefully presented in Section 3.4.1.

3.3 Supervised learning based on binarsity

Consider a random pair $(X, Y) \in \mathbb{R}^p \times \mathscr{Y}$, with unknown distribution $\mathbb{P}$, of which we have n independent samples $D_n = \big((\boldsymbol{X}_{1,\bullet}, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_{n,\bullet}, \boldsymbol{Y}_n)\big)$, and where $\mathscr{Y}$ is the set of responses typically a subset of $\mathbb{R}$. Based on the observations D_n, we have in mind to predict the output Y. This class of problems includes many versions of regression and classification where the conditional distribution of Y given X is completely characterized by the regression function $x \mapsto \mathbb{E}[Y|X = x]$.

Using a statistical learning theory point of view, the goal is to construct a data dependent prediction rule whose risk with respect to a properly chosen loss function, $\ell : \mathscr{Y} \times \mathbb{R} \mapsto \mathbb{R}$, is "close" to the minimal possible risk. We assume that ℓ is convex and continuously differentiable with respect to the second parameter. Typical example of loss functions are the square loss for least square regression, that is $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ with $y \in \mathbb{R}$, and the logistic loss $\ell(y, \hat{y}) = \log(1 + \exp(\hat{y})) - y\hat{y}$, with $y \in \{0, 1\}$. In our approach, we do not assume that $x \mapsto \mathbb{E}[Y|X = x]$ has a particular structure as well as the standard linear approaches in generalized linear regressions. Our estimation procedure for the true regression functions is based on their approximations on the space spanned by the binarized features, namely, $\mathscr{X}^B = \mathrm{span}\{\boldsymbol{X}^B_{i,\bullet}, \text{ for } i = 1, \dots, n\}$. More precisely, using the particularity of $\mathscr{X}^B$, we aim at estimating the true regression functions by a sparse approximation in a given dictionaries of piecewise constant functions.

The statistical performance of a linear predictor $\langle X^B, \theta\rangle$, for some $\theta \in \mathbb{R}^d$ and X^B a binarized copy of X, is measured by the risk

$$R(\theta) = \mathbb{E}[\ell(Y, \langle X^B, \theta\rangle)],$$

If $\hat{\theta}$ is a statistic constructed from the observations D_n, then its risk is given by the conditional expectation

$$R(\hat{\theta}) = \mathbb{E}[\ell(Y, \langle X^B, \hat{\theta}\rangle)|D_n].$$

A natural candidate for the prediction of Y using D_n is the empirical risk minimization procedure, namely any element in $\mathbb{R}^d$ minimizing the empirical risk $R_n(\cdot)$ defined for all $\theta \in \mathbb{R}^d$ by

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(\boldsymbol{Y}_i, \langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle\big).$$

Within this framework, we consider the following penalized optimization problem:

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ R_n(\theta) + \operatorname{bina}_{\hat{w}}(\theta) \right\}. \tag{3.1}$$

An attractive feature of the binaristy penalization is its computational feasibility, since the criterion in (3.1) is convex in θ. Then we can use a convex optimization procedure to compute $\hat{\theta}$ via proximal operator as was seen in Section 3.2.

3.3.1 Linear regression models

We consider the linear regression model $\boldsymbol{Y} = f_0(\boldsymbol{X}) + \boldsymbol{\varepsilon}$, where $f_0(\boldsymbol{X})$ is the true regression function, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^\top$ is an n-dimensional random error, whose components are mean-zero uncorrelated random variables. Without loss of generality we assume the data are centered, so the intercept is not include in the regression model. The estimator of f_0 is defined by $\boldsymbol{X}^B \hat{\theta}$ where $\hat{\theta}$ is the (unique) solution of the following regularized convex problem,

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}^B \theta\|_2^2 + \operatorname{bina}_{\hat{w}}(\theta) \right\}. \tag{3.2}$$

The estimator $\hat{\theta}$ in (3.2) satisfies the following.

Proposition 3.3.1. *Let ε_i be independent random variables and subgaussian with parameter σ_i. Fix $A > 1$ and $d_{\max} = \lfloor A \log p \rfloor$. Consider the estimator $\hat{\theta}$ defined in (3.2). Choose*

$$\hat{w}_{j,k} = \sqrt{2}\sigma_{\max} \sqrt{\frac{d_{\max} \hat{n}_{j,k}}{n}} + \frac{\sqrt{\log 2}}{n}, \tag{3.3}$$

where $\sigma_{\max} = \max_{i=1,\ldots,n} \sigma_i$. Then , with a probability larger than $1 - p^{1-A}$, we have

$$\frac{1}{n} \|f_0(\boldsymbol{X}) - \boldsymbol{X}^B \hat{\theta}\|_2^2 \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|f_0(\boldsymbol{X}) - \boldsymbol{X}^B \theta\|_2^2 + 4 \max_{j=1,\ldots,p} \|\hat{w}_{j,\bullet}\|_\infty \operatorname{bina}(\theta) \right\}. \tag{3.4}$$

The proof of Proposition 3.3.1 is given in Section 3.4.2. In Proposition 3.3.1, the estimator $\hat{\theta}$ satisfies a "slow" oracle inequality, with rate of order $\sqrt{\log d / n}$, which is the expected slow rate involving the b-norm and it holds without any assumption on binarized Gram matrix.

To establish fast binarsity oracle inequalities, we impose a restricted eigenvalue assumption on the binarized matrix $\boldsymbol{X}^B$. Towards this end, we follow the method in Bickel et al. (2009). For all $\theta \in \mathbb{R}^d$, let $J(\theta) = [J_1(\theta), \ldots, J_p(\theta)]$, be the concatenation of the support sets, relative to the total variation penalization, for instance for all $j = 1, \ldots, p$, we define

$$J_j(\theta) = \{k : \theta_{j,k} \neq \theta_{j,k-1}, \text{ for } k = 2, \ldots, d_j\}.$$

Similarly, we set $J^c(\theta) = [J_1^c(\theta), \ldots, J_p^c(\theta)]$, be the complementary of $J(\theta)$. The restricted eigenvalue condition is defined as following.

Assumption 3.3.2. *Let $K = [K_1, \ldots, K_p]$ be a concatenation of index sets. Assume the following condition holds*

$$\kappa_{\boldsymbol{X}^B}(K) = \inf_{u \in \mathbb{R}^d \setminus \{\mathbf{0}_d\} : u \in \mathscr{C}_{\mathrm{bTV},\hat{w}}(K)} \left\{ \frac{\|\boldsymbol{X}^B u\|_2}{\sqrt{n}\|u_K\|_2} \right\} > 0, \tag{3.5}$$

where

$$\mathscr{C}_{\mathrm{bTV},\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j^c}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \leq 3 \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \right\}. \tag{3.6}$$

The set $\mathscr{C}_{\mathrm{bTV},\hat{w}}(K)$ is a cone that involves the bTV-norm and consists of all vectors that have similar support K. Note that the assumption made in Bickel et al. (2009) is slightly stronger but only depends on the cardinality of K, by minimizing with respect to all sets with cardinality equal to one of K. Furthermore, The ℓ_1-norms in Bickel et al. (2009) are now replaced by the weighted TV-norms. The restricted eigenvalue assumption is widely used in the literature and requires somehow that the restriction of the binarized Gram matrix, namely $\Psi_n^B = \frac{1}{n}(\boldsymbol{X}^B)^\top \boldsymbol{X}^B$, in the concatenation of the index sets K is invertible, (when Ψ_n^B is invertible, the condition (3.5) is immediately satisfied with $\kappa_{\boldsymbol{X}^B}(K)$ at least as large as the smallest eigenvalue of Ψ_n^B). We refer to Bickel et al. (2009) for more details on this assumption.

Theorem 3.3.3. *Let Assumption 3.3.2 and those of Proposition 3.3.1 hold. Consider the problem (3.2). Then, with probability at least $1 - p^{1-A}$, for any solution $\hat{\theta}$ of problem (3.2) fulfills the following risk bound*

$$\begin{aligned} &\frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2 \\ &\qquad \leq \inf_{\theta \in \mathscr{H}_1 \times \cdots \times \mathscr{H}_p} \left\{ \frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta\|_2^2 + \frac{608|J(\theta)|}{\kappa^2_{\boldsymbol{X}^B}(J(\theta))} \max_{j=1,\ldots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \right\}. \end{aligned} \tag{3.7}$$

The proof of Theorem 3.3.3 is postponed in Section 3.4.3. An important feature of this inequality is its sharpness, reflected by the fact that the constant in front of the infinimum, often referred to as the leading constant of an oracle inequality, is equal to one. The complexity term depends on both the binarsity and the restricted eigenvalues of the binarized matrix, $\kappa_{\boldsymbol{X}^B}(J(\theta))$. The rate of convergence has the expected shape $\log d/n$.

3.3.2 Generalized linear models

In this section, we focus on the binarsity sparsity to estimate parameters in the generalized linear models introduced in **?**. The random component of a generalized linear model consists of a couple input-output variables (X, Y) where the conditional distribution of Y given $X = x$ is assumed to be from a one parameter exponential dispersion family

$$\exp\left(\frac{y m_0(x) - b(m_0(x))}{a(\phi)} - c(y, \phi) \right)$$

The functions $a(\cdot), b(\cdot)$ and $c(\cdot,\cdot)$ are known, while the natural parameter function $m_0(x)$ is unknown and specifies how the response depends on the feature. Some common examples of generalized linear models are the Poisson regression for count data,

logistic and probit regression for binary data or multinomial regression for categorical data. Often the dispersion parameter function $a(\phi) = 1$ and $c(y, \phi) = c(y)$, giving the natural exponential family of the form $h(y)\exp\big(ym_0(x) - b(m_0(x))\big)$. From now on, we work under the following representation of the conditional distribution

$$f(y; m_0(x)) = \exp\big(ym_0(x) - b(m_0(x))\big).$$

The conditional mean and variance of the $\boldsymbol{Y}$ response are given by

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \dot{b}(m_0(\boldsymbol{X})), \text{ and } m_0(\boldsymbol{X}) = g\big(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}],\big)$$

where the dot denotes differentiation and $\dot{b} = g^{-1}$ is the link function transformation. We consider the empirical risk

$$R_n(m_\theta) = R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(\boldsymbol{Y}_i, \langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle\big),$$

where ℓ is the generalized linear model loss function and is given by

$$\ell\big(\boldsymbol{Y}_i, \langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle\big) = -\boldsymbol{Y}_i\, m_\theta(\boldsymbol{X}_{i,\bullet}) + b(m_\theta(\boldsymbol{X}^B_{i,\bullet})),$$

with $m_\theta(\boldsymbol{X}_{i,\bullet}) = \langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle$. An estimator of m_0 is given by $m_{\hat{\theta}}$, where $\hat{\theta}$ is the solution of the penalized log-likelihood problem

$$\hat{\theta} = \text{argmin}_{\theta\in\mathbb{R}^d}\big\{R_n(\theta) + \text{bina}_{\hat{w}}(\theta)\big\}. \tag{3.8}$$

To evaluate the quality of the estimation, we shall use the excess risk of $m_{\hat{\theta}}$, $R(m_{\hat{\theta}}) - R(m_0)$. Note that

$$\begin{aligned}
R(m_{\hat{\theta}}) - R(m_0) &= \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_{\hat{\theta}}) - R_n(m_0)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\big[b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})) - \dot{b}(m_0(\boldsymbol{X}))m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})\big) \\
&\qquad - \big(b(m_0(\boldsymbol{X}_{i,\bullet})) - \dot{b}(m_0(\boldsymbol{X}))m_0(\boldsymbol{X}_{i,\bullet})\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\big[\dot{b}(m_0(\boldsymbol{X}))m_0(\boldsymbol{X}_{i,\bullet}) - b(m_0(\boldsymbol{X}_{i,\bullet}))\big) \\
&\qquad - \big(\dot{b}(m_0(\boldsymbol{X}))m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}) - b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))\big].
\end{aligned}$$

Now, let us defining the empirical Kullback-Leibler divergence between m_0 and its estimator $\hat{m}$ as follows

$$KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) = \frac{1}{n}\sum_{i=1}^{n} KL\big[f(\boldsymbol{Y}; m_0(\boldsymbol{X}_{i,\bullet})), f(\boldsymbol{Y}; m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))\big].$$

Using this definition, we remark that the excess risk verifies the following.

Lemma 3.3.4. *One has*

$$R(m_{\hat{\theta}}) - R(m_0) = KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})).$$

Proof. Straightforwardly,

$$\begin{aligned}KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) &= KL_n\big[f(\boldsymbol{Y}; m_0(\boldsymbol{X})), f(\boldsymbol{Y}; m_{\hat{\theta}}(\boldsymbol{X}))\big] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathscr{L}(Y|X)}\Big[\log\Big(\frac{f(\boldsymbol{Y}_i; m_0(\boldsymbol{X}_{i,\bullet}))}{f(\boldsymbol{Y}_i; m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))}\Big)\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathscr{L}(Y|X)} \int \Big[\log(f(\boldsymbol{Y}_i; m_0(\boldsymbol{X}_{i,\bullet}))) \\ &\qquad - \log(f(\boldsymbol{Y}_i; m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})))\Big] f(\boldsymbol{Y}_i; m_0(\boldsymbol{X}_{i,\bullet}))d\boldsymbol{Y}_i.\end{aligned}$$

Then

$$\begin{aligned}&KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathscr{L}(Y|X)} \int \Big[\big(\boldsymbol{Y}_i\, m_0(\boldsymbol{X}_{i,\bullet}) - b(m_0(\boldsymbol{X}_{i,\bullet}))\big) \\ &\qquad - \big(\boldsymbol{Y}_i\, m_\theta(\boldsymbol{X}_{i,\bullet}) - b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))\big)\Big] f(\boldsymbol{Y}_i; m_0(\boldsymbol{X}_{i,\bullet}))d\boldsymbol{Y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big[\big(\dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))m_0(\boldsymbol{X}_{i,\bullet}) - b(m_0(\boldsymbol{X}_{i,\bullet}))\big) \\ &\qquad - \big(\dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}) - b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))\big)\Big] \mathbb{E}_{\mathscr{L}(Y|X)} \int f(\boldsymbol{Y}_i; m_0(\boldsymbol{X}_{i,\bullet}))d\boldsymbol{Y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \big[\dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))m_0(\boldsymbol{X}_{i,\bullet}) - b(m_0(\boldsymbol{X}_{i,\bullet}))\big] - \big[\dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}) - b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet}))\big] \\ &= R(m_{\hat{\theta}}) - R(m_0).\end{aligned}$$

■

Assumption 3.3.5. *Assume that there exist two constants $C_n > 0$, $\alpha_i > 0$ and $0 < \xi_n^- \le \xi_n^+$ such that*

$$C_n := \max_{i=1,\dots,n} |m_0(\boldsymbol{X}_{i,\bullet})| < \infty, \tag{3.9}$$

$$\boldsymbol{Y}_i - m_0(\boldsymbol{X}_{i,\bullet}) \text{ is a subgaussian random variable with parameter } \alpha_i, \tag{3.10}$$

$$\xi_n^- \le \max_{i=1,\dots,n} \ddot{b}(m_0(\boldsymbol{X}_{i,\bullet})) \le \xi_n^+. \tag{3.11}$$

In the next table, we give some values of ξ_n^- and ξ_n^+ in the cases of Poisson and Bernoulli distributions.

Distr. \ $b(\cdot)$	$b(z)$	$\dot{b}(z) = g^{-1}(z)$	$\ddot{b}(z)$	ξ_n^-	ξ_n^+
Poisson	$\exp(z)$	$\exp(z)$	$\exp(z)$	$\ddot{b}(-C_n)$	$\ddot{b}(C_n)$
Bernoulli	$\log(1+\exp(z))$	$\frac{\exp(z)}{1+\exp(z)}$	$\frac{\exp(z)}{(1+\exp(z))^2}$	$\ddot{b}(-C_n) \wedge \ddot{b}(C_n)$	$\frac{1}{4}$

In Proposition 3.3.6, the estimator $\hat{\theta}$ satisfies a "slow" oracle inequality, with rate of order $\sqrt{\log d/n}$, which is the expected slow rate involving the b-norm and it holds without any assumption on binarized Gram matrix.

Proposition 3.3.6. *Let Assumption 3.3.5 holds. Fix $A > 1$ and $d_{\max} = \lfloor A \log p \rfloor$. Choose*

$$\hat{w}_{j,k} = \sqrt{2}\alpha_{\max}\sqrt{\frac{d_{\max}\hat{n}_{j,k}}{n}}, \tag{3.12}$$

where $\alpha_{\max} = \max_{i=1,\dots,n} \alpha_i$. *Then, with a probability larger than* $1 - p^{1-A}$, *the estimator* $\hat{\theta}$ *defined in* (3.8) *verifies*

$$KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq \inf_{\theta \in \mathbb{R}^d} \Big\{ KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + 2 \max_{j=1,\dots,p} \|\hat{w}_{j,\bullet}\|_\infty \operatorname{bina}(\theta) \Big\}. \tag{3.13}$$

The proof of Proposition 3.3.6 is given in Section 3.4.2.

We now are interesting in obtaining a non-asymptotic oracle inequality with a fast rate of convergence. To this end, we shall work locally on $B_d(R) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, the ℓ_2-ball of radius $R > 0$ of $\mathbb{R}^d$. This restriction has already been been considered in the case of high-dimensional generalized linear models by van de Geer (2008). Roughly speaking, it means that one can find a set where we can restrict our attention for finding good estimator of m_0. In addition, this restriction on $B_d(R)$, allows us to establish a connection, via the notion of self-concordance see Bach (2010), between the empirical norm $\frac{1}{n}\|\cdot\|_2^2$ and the empirical Kullback divergence $KL_n(\cdot,\cdot)$. Next, we consider the following problem

$$\hat{\theta} = \operatorname{argmin}_{\theta \in B_d(R)} \{R_n(\theta) + \operatorname{bina}_{\hat{w}}(\theta)\}. \tag{3.14}$$

In Theorem 3.3.7, we give a risk bound for the estimator $m_{\hat{\theta}}$ via this empirical Kullback-Leibler divergence.

Theorem 3.3.7. *Let Assumptions 3.3.2 and 3.3.5 be satisfied, and* $\hat{w}_{j,k}$ *as defined in Proposition 3.3.6. Fix* $C_n(R,p,\xi) = \frac{\xi_n^- \psi(-C_n(R,p))}{C_n(R,p)^2}$, $\epsilon > 1/C_n(R,p,\xi)$, *and* $\gamma = 2/\big(\epsilon C_n(R,p,\xi) - 1\big)$, *where* $C_n(R,p) = R\sqrt{p} + C_n$, *and* $\psi(u) = e^u - u - 1$. *Then, with probability greater than* $1 - p^{1-A}$, *for any solution* $\hat{\theta}$ *of problem* (3.14) *fulfills the following risk bound*

$$\begin{aligned} KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq (1+\gamma) \inf_{\theta \in B_d(R) \cap \mathscr{H}_1 \times \cdots \times \mathscr{H}_p} \Big\{ & KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ & + \frac{1216 \epsilon^2 C_n(R,p,\xi)}{(\epsilon C_n(R,p,\xi) - 1) \kappa^2_{\boldsymbol{X}^B}(J(\theta))} |J(\theta)| \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \Big\}. \end{aligned} \tag{3.15}$$

The proof of Theorem 3.3.7 is given in Section 3.4.3. The complexity term depends on both the binarsity and the restricted eigenvalues of the binarized matrix, $\kappa_{\boldsymbol{X}^B}(J(\theta))$. The rate of convergence has the expected shape $\log d/n$.

3.4 Proofs

In this section, we give the proofs of the main results in the paper.

3.4.1 Proof of Proposition 3.2.1 (proximal operator of binarsity)

For a fixed $j = 1, \dots, p$, we aim to prove that $\operatorname{prox}_{\|\cdot\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} + \delta_{H_j}}$ is the composite proximal operators of $\operatorname{prox}_{\|\cdot\|_{\mathrm{bTV},\hat{w}}}$ and $\operatorname{prox}_{\delta_{\mathscr{H}_j}}$, namely

$$\operatorname{prox}_{\|\cdot\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}}(\theta_{j,\bullet}) = \operatorname{prox}_{\delta_{\mathscr{H}_j}}\Big(\operatorname{prox}_{\|\cdot\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})\Big),$$

for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$. Using Theorem 1 in Yu (2013), it is sufficient to show that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$, we have

$$\partial\big(\|\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}\big) \subseteq \partial\big(\|\operatorname{prox}_{\delta_{\mathscr{H}_j}}(\theta_{j,\bullet})\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}\big).$$

Clearly, by the definition of the proximal operator, we have $\text{prox}_{\delta_{\mathscr{H}_j}}(\theta_{j,\bullet}) = \Pi_{\mathscr{H}_j}(\theta_{j,\bullet})$, where $\Pi_{\mathscr{H}_j}(\cdot)$ stands for the projection on the set $\mathscr{H}_j$, i.e., $\Pi_{\mathscr{H}_j}(\theta_{j,\bullet}) = \text{argmin}_{\eta_{j,\bullet}\in\mathscr{H}_j} \|\theta_{j,\bullet} - \eta_{j,\bullet}\|_2$. Besides, we know that $\mathscr{H}_j$ is the hyperplan with normal vector $\mathbb{1}_{j,\bullet}$, then $\mathscr{H}_j$ is the orthogonal subspace of $\text{span}\{\mathbb{1}_{j,\bullet}\}$. Hence

$$\Pi_{\mathscr{H}_j}(\theta_{j,\bullet}) = \Pi_{(\text{span}\{\mathbb{1}_{j,\bullet}\})^\top}(\theta_{j,\bullet}) = \theta_{j,\bullet} - \Pi_{\text{span}\{\mathbb{1}_{j,\bullet}\}}(\theta_{j,\bullet}) = \theta_{j,\bullet} - \bar{\theta}_{j,\bullet}\mathbb{1}_{j,\bullet},$$

where $\bar{\theta}_{j,\bullet} = \frac{1}{d_j}\sum_{k=1}^{d_j}\theta_{j,k}$. Now, let us define the $d_j \times d_j$ matrix D_j by

$$D_j = \begin{bmatrix} 1 & 0 & & & 0 \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ 0 & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d_j} \times \mathbb{R}^{d_j}.$$

We remark that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$,

$$\|\theta_{j,\bullet}\|_{\text{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| = \sum_{k=1}^{d_j} \hat{w}_{j,k}\|D_j\theta_{j,\bullet}\|_1 = \|\hat{w}_{j,\bullet} \odot D_j\theta\|_1.$$

By using subdifferential calculus (see details in the proof of Proposition 3.4.2 below), we have $\partial\big(\|\theta_{j,\bullet}\|_{\text{bTV},\hat{w}_{j,\bullet}}\big) = \partial\big(\|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1\big) = {D_j}^\top \hat{w}_{j,\bullet} \odot \text{sign}(D_j\theta_{j,\bullet})$. Then, it entails that

$$h_{j,\bullet} \in {D_j}^\top \hat{w}_{j,\bullet} \odot \text{sign}(D_j\theta_{j,\bullet}) = {D_j}^\top \hat{w}_{j,\bullet} \odot \text{sign}(D_j(\theta_{j,\bullet} - \bar{\theta}_{j,\bullet}\mathbb{1}_{j,\bullet})).$$

Setting $\beta_{j,\bullet} = \text{prox}_{\|\cdot\|_{\text{bTV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})$ and $\bar{\beta}_{j,\bullet} = \frac{1}{d_j}\sum_{k=1}^{d_j}\beta_{j,k}$ we get

$$\text{prox}_{\|\cdot\|_{\text{bTV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}}(\theta_{j,\bullet}) = \beta_{j,\bullet} - \bar{\beta}_{j,\bullet}\mathbb{1}_{j,\bullet},$$

which gives Algorithm 3.2.1.

□

3.4.2 Proofs of the slow oracle inequalities under binarsity

Proof of Proposition 3.3.1

By the minimizing property of $\hat{\theta}$ it follows that

$$\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}^B\hat{\theta}\|_2^2 + \text{bina}_{\hat{w}}(\hat{\theta}) \leq \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}^B\theta\|_2^2 + \text{bina}_{\hat{w}}(\theta),$$

which, using that $\boldsymbol{Y} = f_0(\boldsymbol{X}) + \boldsymbol{\varepsilon}$, yields

$$\begin{aligned}\frac{1}{2n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2 + \frac{1}{2n}\|\boldsymbol{\varepsilon}\|_2^2 + \frac{1}{n}\langle f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}, \boldsymbol{\varepsilon}\rangle + \text{bina}_{\hat{w}}(\hat{\theta}) \\ \leq \frac{1}{2n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta\|_2^2 + \frac{1}{2n}\|\boldsymbol{\varepsilon}\|_2^2 + \frac{1}{n}\langle f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta, \boldsymbol{\varepsilon}\rangle + \text{bina}_{\hat{w}}(\theta).\end{aligned}$$

Or, equivalently

$$\frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2 \leq \frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta\|_2^2 + \frac{2}{n}\langle\boldsymbol{\varepsilon}, \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle + 2(\text{bina}_{\hat{w}}(\theta) - \text{bina}_{\hat{w}}(\hat{\theta})).$$

So to bound $\frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2$ one must bound $\frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle$.

Let us define the block diagonal matrix $\mathbf{D} = \mathrm{diag}(D_1,\dots,D_p)$. We remark that for all $\theta \in \mathbb{R}^d$,

$$\sum_{j=1}^{p}\sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| = \sum_{j=1}^{p}\sum_{k=1}^{d_j} \hat{w}_{j,k}\|D_j\theta_{j,\bullet}\|_1.$$

Moreover, the matrix D_j is invertible. We denote its inverse T_j and it is defined by the $(d_j \times d_j)$ lower triangular matrix with entries $(T_j)_{r,s} = 0$ if $r < s$ and $(T_j)_{r,s} = 1$ otherwise. We set $\mathbf{T} = \mathrm{diag}(T_1,\dots,T_p)$. Using $\mathbf{TD} = \mathbf{I}_d$, we focus on find out a bound of $\frac{1}{n}\langle (\boldsymbol{X}^B\mathbf{T})\boldsymbol{\varepsilon}, \mathbf{D}(\hat{\theta}-\theta)\rangle$. Let us consider the event

$$\mathscr{U}_n = \bigcap_{j=1}^{p}\bigcap_{k=2}^{d_j} \mathscr{U}_{n,j,k}, \text{ where } \mathscr{U}_{n,j,k} = \Big\{\frac{1}{n}\big|\langle (\boldsymbol{X}^B_{\bullet,j}T_j)_{k,\bullet}, \boldsymbol{\varepsilon}\rangle\big| \le \hat{w}_{j,k}\Big\}.$$

Note that on $\mathscr{U}_n$ one has

$$\begin{aligned}
\frac{2}{n}\langle \boldsymbol{\varepsilon}, \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle &= \frac{2}{n}\sum_{j=1}^{p}\sum_{k=1}^{d_j} \big((\boldsymbol{X}^B_{\bullet,j}T_j)^\top\boldsymbol{\varepsilon}\big)_k\big(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})\big)_k \\
&= \frac{2}{n}\sum_{j=1}^{p}\sum_{k=1}^{d_j} \langle (\boldsymbol{X}^B_{\bullet,j}T_j)_{k,\bullet}, \boldsymbol{\varepsilon}\rangle\big(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})\big)_k \\
&\le 2\sum_{j=1}^{p}\sum_{k=1}^{d_j} \hat{w}_{j,k}\big|\big(D_j(\theta_{j,\bullet} - \hat{\theta}_{j,\bullet})\big)_k\big| \\
&\le 2\sum_{j=1}^{p}\Big(\|\theta_{j,\bullet} - \hat{\theta}_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}(\theta_{j,\bullet} - \hat{\theta}_{j,\bullet})\Big) \\
&= 2\,\mathrm{bina}_{\hat{w}}(\theta - \hat{\theta}).
\end{aligned}$$

Putting things together, on $\mathscr{U}_n$

$$\frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2 \le \frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta\|_2^2 + 2\big(\mathrm{bina}_{\hat{w}}(\theta-\hat{\theta}) + \mathrm{bina}_{\hat{w}}(\theta) - \mathrm{bina}_{\hat{w}}(\hat{\theta})\big).$$

Next, we prove that binarsity is a subadditive penalization.

Lemma 3.4.1. *For all $\theta, \theta' \in \mathbb{R}^d$, one has*

$$\mathrm{bina}(\theta+\theta') \le \mathrm{bina}(\theta) + \mathrm{bina}(\theta'), \textit{and} \tag{3.16}$$

$$\mathrm{bina}(-\theta) \le \mathrm{bina}(\theta). \tag{3.17}$$

Proof. The hyperplan $\mathscr{H}_j$ is a convex cone, then the indicator function $\delta_{\mathscr{H}_j}$ is sublinear (positively homogeneous + subadditive, see **?**). Furthermore, the total variation penalization satisfies triangular inequality, which gives (3.16).

To prove (3.17), we use the fact that $\delta_{\mathscr{H}_j}(\theta_{j,\bullet}) + \delta_{\mathscr{H}_j}(-\theta_{j,\bullet}) \ge 0$, then we obtain

$$\begin{aligned}
\mathrm{bina}(-\theta) &= \sum_{j=1}^{p}\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV}} + \delta_{\mathscr{H}_j}(-\theta)\Big) \\
&\le \sum_{j=1}^{p}\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV}} - \delta_{\mathscr{H}_j}(\theta))\Big) \\
&\le \sum_{j=1}^{p}\Big(\|\theta_{j,\bullet}\|_{\mathrm{TV}} + \delta_{\mathscr{H}_j}(\theta))\Big).
\end{aligned}$$

■

By Lemma 3.4.1, we get

$$\begin{aligned}\frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\hat{\theta}\|_2^2 &\leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+4\,\mathrm{bina}_{\hat{w}}(\theta)\\ &\leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+4\max_{j=1,\dots,p}\|\hat{w}_{j,\bullet}\|_\infty\,\mathrm{bina}(\theta)\end{aligned}$$

Now, using an elementary bound on an elementary bound on the tails of independent subgaussian random variables given by Hoeffding inequality, and the choice of the weights $\hat{w}_{j,k}$ for all $j\in\{1,\dots,p\}$, and $k\in\{2,\dots,d_j\}$ given by (3.3), we find that the probability of the complementary event $\mathscr{U}_n^c$ is equal to p^{1-A}. This concludes the proof.

□

Proof of Proposition 3.3.6

Recall that for all $\theta\in\mathbb{R}^d$,

$$R_n(\theta)=\frac{1}{n}\sum_{i=1}^n b(m_\theta(\boldsymbol{X}^B_{i,\bullet}))-\frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i m_\theta(\boldsymbol{X}^B_{i,\bullet}),$$

and

$$\hat{\theta}=\mathrm{argmin}_{\theta\in\mathbb{R}^d}\left\{R_n(\theta)+\mathrm{bina}_{\hat{w}}(\theta)\right\}.$$

First, we observe that

$$R(\theta)=\mathbb{E}[R_n(\theta)]=R_n(\theta)+\frac{1}{n}\langle \boldsymbol{Y}-m_0(\boldsymbol{X}),m_\theta(\boldsymbol{X}^B)\rangle.$$

Using the mere definition of $\hat{\theta}$, we have

$$R(\hat{\theta})-\frac{1}{n}\langle \boldsymbol{Y}-m_0(\boldsymbol{X}),\boldsymbol{X}^B\hat{\theta}\rangle+\mathrm{bina}_{\hat{w}}(\hat{\theta})\leq R(\theta)-\frac{1}{n}\langle \boldsymbol{Y}-m_0(\boldsymbol{X}),\boldsymbol{X}^B\hat{\theta}\rangle+\mathrm{bina}_{\hat{w}}(\theta).$$

Hence we get

$$\begin{aligned}KL_n(m_0(\boldsymbol{X}),m_{\hat{\theta}}(\boldsymbol{X}))\leq{}& KL_n(m_0(\boldsymbol{X}),m_\theta(\boldsymbol{X}))\\ &+\frac{1}{n}\langle \boldsymbol{Y}-m_0(\boldsymbol{X}),\boldsymbol{X}^B(\hat{\theta}-\theta)\rangle+\mathrm{bina}_{\hat{w}}(\theta)-\mathrm{bina}_{\hat{w}}(\hat{\theta}).\end{aligned}$$

Then, consider

$$\mathscr{V}_n=\bigcap_{j=1}^p\bigcap_{k=2}^{d_j}\mathscr{V}_{n,j,k},\ \text{where}\ \mathscr{V}_{n,j,k}=\left\{\frac{1}{n}\left|\langle\left(\boldsymbol{X}^B_{\bullet,j}T_j\right)_{k,\bullet},\boldsymbol{Y}-m_0(\boldsymbol{X})\rangle\right|\leq\hat{w}_{j,k}\right\}.$$

Thus, on $\mathscr{V}_n$ we have

$$\begin{aligned}KL_n(m_0(\boldsymbol{X}),m_{\hat{\theta}}(\boldsymbol{X}))&\leq KL_n(m_0(\boldsymbol{X}),m_\theta(\boldsymbol{X}))+\mathrm{bina}_{\hat{w}}(\theta-\hat{\theta})+\mathrm{bina}_{\hat{w}}(\theta)-\mathrm{bina}_{\hat{w}}(\hat{\theta})\\ &\leq KL_n(m_0(\boldsymbol{X}),m_\theta(\boldsymbol{X}))+2\,\mathrm{bina}_{\hat{w}}(\theta)\\ &\leq KL_n(m_0(\boldsymbol{X}),m_\theta(\boldsymbol{X}))+2\max_{j=1,\dots,p}\|\hat{w}_{j,\bullet}\|_\infty\,\mathrm{bina}(\theta).\end{aligned}$$

To finish the proof, we prove that the probability of the complementary event $\mathscr{V}_n^c$ is equal to $0.25e^{-A}$ using an elementary bound on the tails of subgaussian random variables given by Hoeffding inequality, and the choice of the weights $\hat{w}_{j,k}$ for all $j\in\{1,\dots,p\}$, and $k\in\{2,\dots,d_j\}$, given by (3.12). This gives the desired result.

□

3.4.3 Proofs of the fast oracle inequalities under binaristy

Optimality conditions

To characterize the solution of the problem (3.1), the following result can be sraightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions (see Boyd and Vandenberghe (2004)) for a convex optimization.

Proposition 3.4.2. *A vector $\hat{\theta} = (\hat{\theta}_{1,\bullet}^\top \cdots \hat{\theta}_{p,\bullet}^\top)^\top \in \mathbb{R}^d$, is an optimum of the objective function in (3.1) if and only if there exists a sequence of subgradients $\hat{h} = (\hat{h}_{j,\bullet})_{j=1,\ldots,p} \in \partial\big(\|\hat{\theta}\|_{\mathrm{bTV},\hat{w}}\big)$ and $\hat{g} = (\hat{g}_{j,\bullet})_{j=1,\ldots,p} \in \partial\big(\delta_{\mathscr{H}_j}(\hat{\theta}_{j,\bullet})\big)_{j=1,\ldots,p}$, such that*

$$\nabla R_n(\hat{\theta}_{j,\bullet}) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} = \mathbf{0}_{d_j},$$

where

$$\hat{h}_{j,\bullet} \begin{cases} = D_j^\top\big(\hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\hat{\theta}_{j,\bullet})\big), & \text{if } j \in J(\hat{\theta}), \\ \in D_j^\top\big(\hat{w}_{j,\bullet} \odot [-1,+1]^{d_j}\big), & \text{if } j \in J^c(\hat{\theta}) \end{cases}$$

and where $J(\hat{\theta})$ is the active set of $\hat{\theta}$. The subgradient $\hat{g}_{j,\bullet}$ belongs to

$$\partial\big(\delta_{\mathscr{H}_j}(\hat{\theta}_{j,\bullet})\big) = \big\{\mu_{j,\bullet} \in \mathbb{R}^{d_j} : \langle \mu_{j,\bullet}, \theta_{j,\bullet}\rangle \le \langle \mu_{j,\bullet}, \hat{\theta}_{j,\bullet}\rangle, \text{ for all } \theta_{j,\bullet} \in \mathscr{H}_j\big\}. \tag{3.18}$$

If we consider the linear regression model, see problem (3.2), we have

$$\frac{1}{n}\big(\boldsymbol{X}^B_{\bullet,j}\big)^\top(\boldsymbol{X}^B\hat{\theta} - \boldsymbol{Y}) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} = \mathbf{0}_{d_j}. \tag{3.19}$$

Additionally, for the generalized linear model, see problem (3.8), we have

$$\frac{1}{n}\big(\boldsymbol{X}^B_{\bullet,j}\big)^\top\big(\dot{b}(\boldsymbol{X}^B\hat{\theta}) - \boldsymbol{Y}\big) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} + \hat{f}_{j,\bullet} = \mathbf{0}_{d_j}, \tag{3.20}$$

where $\hat{f} = (\hat{f}_{j,\bullet})_{j=1,\ldots,p}$ belongs to the normal cone to the ball $B_d(R)$.

Proof of Proposition 3.4.2

We denote by $\partial(\phi)$ the subdifferential mapping of a convex functional ϕ. The function $\theta \mapsto R_n(\theta)$ is differentiable, so the subdifferential of $R_n(\cdot) + \mathrm{bina}_{\hat{w}}(\cdot)$ at a point $\theta = (\theta_{j,\bullet})_{j=1,\ldots,p} \in \mathbb{R}^d$ is given by

$$\partial\big(R_n(\theta) + \mathrm{bina}_{\hat{w}}(\theta)\big) = \big\{\nabla R_n(\theta)\big\} + \partial\Big(\sum_{j=1}^p \|\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} + \delta_{\mathscr{H}_j}(\theta_{j,\bullet})\Big),$$

where $\nabla R_n(\theta) = \Big(\frac{\partial(R_n(\theta))}{\partial(\theta_{1,\bullet})}, \ldots, \frac{\partial(R_n(\theta))}{\partial(\theta_{p,\bullet})}\Big)^\top$ and

$$\partial\big(\mathrm{bina}_{\hat{w}}(\theta)\big) = \Big(\partial\big(\|\theta_{1,\bullet}\|_{\mathrm{bTV},\hat{w}_{1,\bullet}}\big) + \partial\big(\delta_{\mathscr{H}_1}(\theta_{1,\bullet})\big), \ldots, \partial\big(\|\theta_{p,\bullet}\|_{\mathrm{bTV},\hat{w}_{p,\bullet}}\big) + \partial\big(\delta_{\mathscr{H}_p}(\theta_{p,\bullet})\big)\Big)^\top.$$

For all $j = 1,\ldots,p$, we have $\|\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} = \|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1$. Then, by applying some properties of the subdifferential calculus we get,

$$\forall j = 1,\ldots,p,\ \partial\big(\|\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}\big) = \begin{cases} D_j^\top \mathrm{sign}(\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}), & \text{if } D_j\theta \neq \mathbf{0}_{d_j} \\ D_j^\top\big(\hat{w}_{j,\bullet} \odot [-1,+1]^{d_j}\big), & \text{otherwise}, \end{cases} \tag{3.21}$$

where $[-1,+1]^{d_j} = \{[-1,+1],\ldots,[-1,+1]\}$.

Now, $\hat{\theta} = (\hat{\theta}_{1,\bullet}^\top \cdots \hat{\theta}_{p,\bullet}^\top)^\top$ is an optimum of the problem (3.2), if and only if for all $j = 1,\ldots,p$, $\mathbf{0}_{d_j} \in \nabla R_n(\hat{\theta}_{j,\bullet}) + \partial\big(\|\hat{\theta}_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}\big) + \partial\big(\delta_{\mathscr{H}_j}(\hat{\theta}_{j,\bullet})\big)$. Using (3.21) and $\frac{\partial(R_n(\theta))}{\partial(\theta_{j,\bullet})} = \frac{1}{n}\big(\boldsymbol{X}^B_{\bullet,j}\big)^\top(\boldsymbol{X}^B\hat{\theta} - \boldsymbol{Y})$, then (3.19) holds.

For generalized linear models, we rewrite the problem (3.8) as follows

$$\hat{\theta} = \mathrm{argmin}_{\theta\in\mathbb{R}^d}\left\{R_n(\theta) + \mathrm{bina}_{\hat{w}}(\theta) + \delta_{B_d(R)}(\theta)\right\},$$

where $\delta_{B_d(R)}$ is the indicator function to the $B_d(R)$. Similarly, $\hat{\theta} = (\hat{\theta}_{1,\bullet}^\top \cdots \hat{\theta}_{p,\bullet}^\top)^\top$, is an optimum of the problem (3.14) if and only if $\mathbf{0}_d \in \nabla R_n(\hat{\theta}) + \partial\big(\|\hat{\theta}\|_{\mathrm{bTV},\hat{w}}\big) + \partial\big(\delta_{B_d(R)}(\hat{\theta})\big)$. Recall that the subdifferential of $\delta_{B_d(R)}(\cdot)$ is the normal cone to $B_d(R)$, i.e.,

$$\partial\big(\delta_{B_d(R)}(\hat{\theta})\big) = \big\{\eta\in\mathbb{R}^d : \langle\eta,\theta\rangle \le \langle\eta,\hat{\theta}\rangle, \text{ for all } \theta\in B_d(R)\big\}. \tag{3.22}$$

Straightforwardly,

$$\begin{aligned}\frac{\partial(R_n(\theta))}{\partial(\theta_{j,\bullet})} &= \frac{1}{n}\big(\boldsymbol{X}^B_{\bullet,j}\big)^\top\big(\dot{b}(\boldsymbol{X}^B\hat{\theta}) - \boldsymbol{Y}\big)\\ &= \frac{1}{n}(\boldsymbol{X}^B_{\bullet,j})^\top\big(\dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{Y}\big).\end{aligned} \tag{3.23}$$

Equalities (3.23) and (3.22) give the Equation (3.20) in Proposition 3.4.2.

□

Compatibility condition for the matrix T

To establish fast oracle inequalities we need, in addition to Assumption 3.3.2, the following result which gives a compatibility condition satisfied by the matrix $\mathbf{T}$. To this end, for any concatenation of subsets $K = [K_1,\ldots,K_p]$, we set for all $j = 1,\ldots,p$,

$$K_j = \{\tau_j^1,\ldots,\tau_j^{b_j}\} \subset \{1,\ldots,d_j\}, \text{ with the convention that } \tau_j^0 = 0, \text{ and } \tau_j^{b_j+1} = d_j + 1. \tag{3.24}$$

Lemma 3.4.3. *Let $\gamma\in\mathbb{R}_+^d$ be a given vector of "weights". For every $K = [K_1,\ldots,K_p]$, such that for all $j = 1,\ldots,p, K_j$ is given by (3.24), and for every $u\in\mathbb{R}^d\setminus\{\mathbf{0}_d\}$, we have*

$$\frac{\|\mathbf{T}u\|_2}{\Big|\|u_K\odot\gamma_K\|_1 - \|u_{K^c}\odot\gamma_{K^c}\|_1\Big|} \ge \kappa_{\mathbf{T},\gamma}(K), \tag{3.25}$$

where

$$\kappa_{\mathbf{T},\gamma}(K) = \Big\{32\sum_{j=1}^p\sum_{k=1}^{d_j}|\gamma_{j,k+1}-\gamma_{j,k}|^2 + (b_j+1)\|\gamma_{j,\bullet}\|_\infty^2\Delta_{\min,K_j}^{-1}\Big\}^{-1/2},$$

and $\Delta_{\min,K_j} = \min_{r\in[b^j]}|\tau_j^{r_j} - \tau_j^{r_j-1}|$.

Proof of Lemma 3.4.3

Using Proposition 3 in Dalalyan et al. (2014), we obtain

$$\begin{aligned}&\|u_K\odot\gamma_K\|_1 - \|u_{K^c}\odot\gamma_{K^c}\|_1\\ &\quad= \sum_{j=1}^p \|u_{K_j}\odot\gamma_{K_j}\|_1 - \|u_{K_j^c}\odot\gamma_{K_j^c}\|_1\\ &\quad\le \sum_{j=1}^p 4\|T_j u_{j,\bullet}\|_2\Big\{2\sum_{k=1}^{d_j}|\gamma_{j,k+1}-\gamma_{j,k}|^2 + 2(b_j+1)\|\gamma_{j,\bullet}\|_\infty^2\Delta_{\min,K_j}^{-1}\Big\}^{1/2}.\end{aligned}$$

Applying Hölder's inequality for the right hand side on the last inequality, we get

$$\begin{aligned}&\|u_K \odot \gamma_K\|_1 - \|u_{K^c} \odot \gamma_{K^c}\|_1 \\ &\qquad \leq \|\mathbf{T}u\|_2 \Big\{ 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + (b_j+1)\|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1} \Big\}^{1/2}.\end{aligned}$$

This complete the proof. □

Compatibility condition of $\boldsymbol{X}^B\mathbf{T}$

Now, using Assumtion 3.3.2 and Lemma 3.4.3, we arrive at establishing a compatibility condition satisfied by the product of matrices $\boldsymbol{X}^B\mathbf{T}$.

Lemma 3.4.4. *Let Assumption 3.3.2 holds. Let $\gamma \in \mathbb{R}_+^d$, be a given vector of "weights", and $K = [K_1, \ldots, K_p]$, such that for all $j = 1, \ldots, p, K_j$ is given by (3.24). Then, we have*

$$\inf_{u \in \mathbb{R}^d \setminus \{\mathbf{0}_d\} : u \in \mathscr{C}_{1,\hat{w}}(K)} \left\{ \frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\big|\|u_K \odot \gamma_K\|_1 - \|u_{K^c} \odot \gamma_{K^c}\|_1\big|} \right\} \geq \kappa_{\mathbf{T},\gamma}(K)\kappa_{\boldsymbol{X}^B}(K), \tag{3.26}$$

where

$$\mathscr{C}_{1,\hat{w}}(K) = \Big\{ u \in \mathbb{R}^d : \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j^c}\|_{1,\hat{w}_{j,\bullet}} \leq 3 \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j}\|_{1,\hat{w}_{j,\bullet}} \Big\}, \tag{3.27}$$

with $\|\cdot\|_{1,a}$ denotes the weighted ℓ_1-norm.

Proof of Lemma 3.4.4

By Lemma 3.4.3, we have that

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\big|\|u_K \odot \gamma_K\|_1 - \|u_{K^c} \odot \gamma_{K^c}\|_1\big|} \geq \kappa_{\mathbf{T},\gamma}(K) \frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\|\mathbf{T}u\|_2}.$$

Now, we note that if $u \in \mathscr{C}_{1,\hat{w}}(K)$ then $\mathbf{T}u \in \mathscr{C}_{\mathrm{bTV},\hat{w}}(K)$. Hence, by Assumption 3.3.2, we get

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n}\big|\|u_K \odot \gamma_K\|_1 - \|u_{K^c} \odot \gamma_{K^c}\|_1\big|} \geq \kappa_{\mathbf{T},\gamma}(K)\kappa_{\boldsymbol{X}^B}(K).$$

□

Proof of Theorem 3.3.3

Recall that for all $\theta \in \mathbb{R}^d$, $R_n(\theta) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}^B\theta\|_2^2$, and

$$\hat{\theta} = \mathrm{argmin}_{\theta \in \mathbb{R}^d} \Big\{ R_n(\theta) + \mathrm{bina}_{\hat{w}}(\theta) \Big\}. \tag{3.28}$$

By Proposition 3.4.2, Equation (3.28) means that there is $\hat{h} = (\hat{h}_{j,\bullet})_{j=1,\ldots,p} \in \partial\big(\|\hat{\theta}\|_{\mathrm{bTV},\hat{w}}\big)$ and $\hat{g} = (\hat{g}_{j,\bullet})_{j=1,\ldots,p} \in \partial\big(\delta_{\mathscr{H}_j}(\hat{\theta}_{j,\bullet})\big)$ such that

$$\langle \frac{1}{n}\big(\boldsymbol{X}^B\big)^\top(\boldsymbol{X}^B\hat{\theta} - \boldsymbol{Y}) + \hat{h} + \hat{g}, \hat{\theta} - \theta \rangle = 0, \text{ for all } \theta \in \mathbb{R}^d. \tag{3.29}$$

Equation (3.29) can be written in the following way:

$$\langle \frac{1}{n}\big(\boldsymbol{X}^B\big)^\top(\boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X})) - \frac{1}{n}(\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X})), \hat{\theta} - \theta\rangle + \langle \hat{h}, \hat{\theta} - \theta\rangle + \langle \hat{g}, \hat{\theta} - \theta\rangle = 0. \tag{3.30}$$

Now, using the fact that the subdifferential mapping is monotone (this is an immediate consequence of its definition, see Rockafellar (1970), to say that $\langle \hat{h} - h, \hat{\theta} - \theta\rangle \geq 0$, for any $h \in \partial\big(\|\theta\|_{\mathrm{bTV},\hat{w}}\big)$ entails that

$$\frac{2}{n}\langle \boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X}), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \leq \frac{2}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}), \hat{\theta} - \theta\rangle - 2\langle h, \hat{\theta} - \theta\rangle - \langle \hat{g}, \hat{\theta} - \theta\rangle.$$

Furthermore, for any $\theta = (\theta_{j,\bullet}^\top \cdots \theta_{p,\bullet}^\top)^\top \in \mathscr{H}_1 \times \cdots \times \mathscr{H}_p$ we have $-\langle \hat{g}_{j,\bullet}, \hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\rangle \leq 0$. It implies

$$\frac{2}{n}\langle \boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X}), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \leq \frac{2}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}), \hat{\theta} - \theta\rangle - 2\langle h, \hat{\theta} - \theta\rangle.$$

According to Al-Kashi formula, we have

$$\begin{aligned}\frac{2}{n}&\langle \boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X}), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \\ &= \frac{1}{n}\|\boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X})\|_2^2 + \frac{1}{n}\|\boldsymbol{X}^B(\hat{\theta} - \theta)\|_2^2 - \frac{1}{n}\|\boldsymbol{X}^B\theta - f_0(\boldsymbol{X})\|_2^2.\end{aligned} \tag{3.31}$$

Then

$$\begin{aligned}\frac{1}{n}&\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\hat{\theta}\|_2^2 + \frac{1}{n}\|\boldsymbol{X}^B(\hat{\theta} - \theta)\|_2^2 \\ &\leq \frac{1}{n}\|f_0(\boldsymbol{X}) - \boldsymbol{X}^B\theta\|_2^2 + \frac{2}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}), \hat{\theta} - \theta\rangle - 2\langle h, \hat{\theta} - \theta\rangle.\end{aligned} \tag{3.32}$$

If $\langle \boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X}), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle < 0$, we have $\frac{1}{n}\|\boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X})\|_2^2 < \frac{1}{n}\|\boldsymbol{X}^B\theta - f_0(\boldsymbol{X})\|_2^2$, which yields Theorem 3.3.3, so we assume that $\langle \boldsymbol{X}^B\hat{\theta} - f_0(\boldsymbol{X}), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \geq 0$. In this case, we obtain

$$\frac{2}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}), \hat{\theta} - \theta\rangle - 2\langle h, \hat{\theta} - \theta\rangle \geq 0.$$

In one hand, using that the matrix $\mathbf{D}$ and its inverse is given by $\mathbf{D}^{-1} = \mathbf{T}$, we have

$$\begin{aligned}\frac{2}{n}&\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}), \hat{\theta} - \theta\rangle \\ &= \frac{2}{n}\langle (\boldsymbol{X}^B\mathbf{T})^\top(Y - f_0(\boldsymbol{X}), \mathbf{D}(\hat{\theta} - \theta)\rangle \\ &= \frac{2}{n}\sum_{j=1}^p \langle (\boldsymbol{X}^B_{\bullet,j}T_j)^\top(\boldsymbol{Y} - f_0(\boldsymbol{X})), D_j\big(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\big)\rangle \\ &= \frac{2}{n}\sum_{j=1}^p\sum_{k=1}^{d_j}\Big((\boldsymbol{X}^B_{\bullet,j})^\top(\boldsymbol{Y} - f_0(\boldsymbol{X}))\Big)_k\Big(D_j\big(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\big)\Big)_k \\ &\leq \frac{2}{n}\sum_{j=1}^p\sum_{k=1}^{d_j}\Big|\langle (\boldsymbol{X}^B_{\bullet,j}T_j)_{k,\bullet}, \boldsymbol{Y} - f_0(\boldsymbol{X})\rangle\Big|\Big|\Big(D_j\big(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\big)\Big)_k\Big|.\end{aligned} \tag{3.33}$$

Let us define the event

$$\mathscr{I}_n = \bigcap_{j=1}^p\bigcap_{k=2}^{d_j}\mathscr{I}_{n,j,k}, \text{ where } \mathscr{I}_{n,j,k} = \Big\{\frac{1}{n}\Big|\langle (\boldsymbol{X}^B_{\bullet,j}T_j)_{k,\bullet}, \boldsymbol{Y} - f_0(\boldsymbol{X})\rangle\Big| \leq \frac{\hat{w}_{j,k}}{2}\Big\}. \tag{3.34}$$

Here the vector $\left(\boldsymbol{X}^B_{\bullet,j}T_j\right)_{k,\bullet}$ denotes the k^{th} column of the $(n \times d_j)$ matrix $\boldsymbol{X}^B_{\bullet,j}T_j$. Then, on $\mathcal{I}_n$, we have that

$$\begin{aligned}\frac{2}{n}\langle(\boldsymbol{X}^B)^\top(\boldsymbol{Y}-f_0(\boldsymbol{X}),\hat{\theta}-\theta\rangle &\leq \sum_{j=1}^{p}\sum_{k=1}^{d_j}\hat{w}_{j,k}\left|\left(D_j\left(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet}\right)\right)_k\right| \\ &= \sum_{j=1}^{p}\|\hat{\theta}_{j,\bullet}-\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\ &= \sum_{j=1}^{p}\|\hat{w}_{j,\bullet}\odot D_j\left(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet}\right)\|_1. \end{aligned} \tag{3.35}$$

In another hand, from the definition of the subgradient $(h_{j,\bullet})_{j=1,\dots,p} \in \partial\left(\|\theta\|_{\mathrm{bTV},\hat{w}}\right)$ for a fixed $\theta \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$, we can choose h such that

$$\begin{aligned} h_{j,\bullet} &= \left(D_j^\top\left(\hat{w}_{j,\bullet}\odot \mathrm{sign}(D_j\theta_{j,\bullet})\right)\right)_{k\in\{1,\dots,J_j(\theta)\}}, \\ h_{j,\bullet} &= \left(D_j^\top\left(\hat{w}_{j,\bullet}\odot \mathrm{sign}\left(D_j\hat{\theta}_{j,\bullet}\right)\right)\right)_{k\in\{1,\dots,J_j^c(\theta)\}} = \left(D_j^\top\left(\hat{w}_{j,\bullet}\odot \mathrm{sign}\left(D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})\right)\right)\right)_{k\in\{1,\dots,J_j^c(\theta)\}}. \end{aligned}$$

This gives

$$\begin{aligned} &-2\langle h,\hat{\theta}-\theta\rangle \\ &\quad= -2\sum_{j=1}^{p}\langle h_{j,\bullet},\hat{\theta}_{j,\bullet}-\theta_{j,\bullet}\rangle \\ &\quad= 2\sum_{j=1}^{p}\langle(-h_{j,\bullet})_{J_j(\theta)},(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\rangle - 2\sum_{j=1}^{p}\langle(h_{j,\bullet})_{J_j^c(\theta)},(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\rangle \\ &\quad= 2\sum_{j=1}^{p}\langle(-\hat{w}_{j,\bullet}\odot\mathrm{sign}(D_j\theta_{j,\bullet}))_{J_j(\theta)},D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\rangle \\ &\qquad\qquad - 2\sum_{j=1}^{p}\langle(\hat{w}_{j,\bullet}\odot\mathrm{sign}\left(D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})\right))_{J_j^c(\theta)},D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\rangle. \end{aligned}$$

Using a triangle inequality and the fact that $\langle\mathrm{sign}(x),x\rangle = \|x\|_1$, imply that

$$\begin{aligned} &-2\langle h,\hat{\theta}-\theta\rangle \\ &\quad\leq 2\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_1 \\ &\qquad\qquad - 2\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j^c(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_1 \\ &\quad= 2\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} - 2\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}. \end{aligned} \tag{3.36}$$

Together Inequalities (3.35) and (3.36) with the fact that on $\mathcal{I}_n$, we have established that

$$\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \leq 3\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}, \tag{3.37}$$

also

$$\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j^c(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_1 \leq 3\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_1. \tag{3.38}$$

This means, (see (3.6) and (3.27)), that

$$\hat{\theta}-\theta \in \mathscr{C}_{\mathrm{bTV},\hat{w}}(J(\theta)), \text{ and } \mathbf{D}(\hat{\theta}-\theta) \in \mathscr{C}_{1,\hat{w}}(J(\theta)). \tag{3.39}$$

Now, returning to (3.32), we arrive at

$$\begin{aligned}
&\frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\hat{\theta}\|_2^2+\frac{1}{n}\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2^2 \\
&\leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+\sum_{j=1}^{p}\|\hat{\theta}_{j,\bullet}-\theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\
&\qquad +2\sum_{j=1}^{p}\|((\hat{\theta})_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}-2\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\
&\leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}+\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\
&\qquad +2\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}-2\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\
&\leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+3\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}-\sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}.
\end{aligned}$$

Equivalently

$$\begin{aligned}
&\frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\hat{\theta}\|_2^2+\frac{1}{n}\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2^2 \\
&\qquad \leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+3\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_1 \\
&\qquad\qquad -\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j^c(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j^c(\theta)}\|_1 \\
&\qquad \leq \frac{1}{n}\|f_0(\boldsymbol{X})-\boldsymbol{X}^B\theta\|_2^2+3\sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\odot D_j(\hat{\theta}_{j,\bullet}-\theta_{j,\bullet})_{J_j(\theta)}\|_1.
\end{aligned} \tag{3.40}$$

In inequality (3.40), by using (3.39) and Lemma 3.4.4, we see that

$$\frac{1}{n}\|\boldsymbol{X}^B\hat{\theta}-f_0(\boldsymbol{X})\|_2^2+\frac{1}{n}\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2^2 \leq \frac{1}{n}\|\boldsymbol{X}^B\theta-f_0(\boldsymbol{X})\|_2^2+2\frac{\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2}{\sqrt{n}\,\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))},$$

where $\hat{\gamma}=(\hat{\gamma}_{1,\bullet}^\top,\cdots\hat{\gamma}_{p,\bullet}^\top)^\top$ such that

$$\forall j=1,\ldots,p,\ \hat{\gamma}_{j,k}=\begin{cases} 3\hat{w}_{j,k}, & \text{if } k\in J_j(\theta),\\ 0, & \text{if } k\in J_j^c(\theta),\end{cases}$$

and

$$\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))=\Big\{32\sum_{j=1}^{p}\sum_{k=1}^{d_j}|\hat{\gamma}_{j,k+1}-\hat{\gamma}_{j,k}|^2+2|J_j(\theta)|\|\hat{\gamma}_{j,\bullet}\|_\infty^2\Delta_{\min,J_j(\theta)}^{-1}\Big\}^{-1/2}.$$

Using the fact that $2uv\leq u^2+v^2$, it implies that

$$\frac{1}{n}\|\boldsymbol{X}^B\hat{\theta}-f_0(\boldsymbol{X})\|_2^2 \leq \frac{1}{n}\|\boldsymbol{X}^B\theta-f_0(\boldsymbol{X})\|_2^2+\frac{1}{\kappa_{\mathbf{T},\hat{\gamma}}^2(J(\theta))}\times\frac{1}{\kappa_{\boldsymbol{X}^B}^2(J(\theta))}.$$

To finish the proof, it sufficient to find an upper bound for $\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))}$ which is given by

$$\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))} = 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 + 2|J_j(\theta)| \|\hat{\gamma}_{j,\bullet}\|^2_\infty \Delta^{-1}_{\min,J_j(\theta)}.$$

Note that $\|\hat{\gamma}_{j,\bullet}\|_\infty \leq 3\|\hat{w}_{j,\bullet}\|_\infty$. We write the set $J_j(\theta) = \{k_j^1,...,k_j^{|J_j(\theta)|}\}$ and we set $B_r = [\![k_j^{r-1}, k_j^r[\![= \{k_j^{r-1}, k_j^{r-1}+1, \ldots, k_j^r - 1\}$ for $r \in \{1,\ldots,|J_j(\theta)|+1\}$ with the convention that $k_j^0 = 1$, and $k_j^{|J_j(\theta)|+1} = d_j + 1$. Then

$$\begin{aligned}
\sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 &= \sum_{r=1}^{|J_j(\theta)|+1} \sum_{k \in B_r} |\gamma_{j,k+1} - \gamma_{j,k}|^2 \\
&= \sum_{r=1}^{|J_j(\theta)|+1} \left\{ |\hat{\gamma}_{j,k_j^{r-1}+1} - \hat{\gamma}_{j,k_j^{r-1}}|^2 + |\hat{\gamma}_{j,k_j^r} - \hat{\gamma}_{j,k_j^r-1}|^2 \right\} \\
&= \sum_{r=1}^{|J_j(\theta)|+1} \left\{ \hat{\gamma}^2_{j,k_j^{r-1}} + \hat{\gamma}^2_{j,k_j^r} \right\} \\
&= \sum_{r=1}^{|J_j(\theta)|} 2\hat{\gamma}^2_{j,k_j^r} \\
&\leq 18|J_j(\theta)| \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))} &\leq 32 \sum_{j=1}^{p} \left\{ 18|J_j(\theta)| \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty \right\} + 18|J_j(\theta)| \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty \Delta^{-1}_{\min,J_j(\theta)} \\
&\leq 32 \sum_{j=1}^{p} \left\{ 18 + \frac{1}{\Delta_{\min,J_j(\theta)}} \right\} |J_j(\theta)| \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty \\
&\leq 608|J(\theta)| \max_{j=1,\ldots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.
\end{aligned}$$

Finally, using an elementary bound on an elementary bound on the tails of subgaussian random variables given by Hoeffding inequality, and the choice of the weights $\hat{w}_{j,k}$ for all $j \in \{1,\ldots,p\}$, and $k \in \{2,\ldots,d_j\}$ by (3.3), we find that the probability of the complementary event $\mathcal{X}_n^c$ is equal to to p^{1-A}. This concludes the proof.

□

Proof of fast oracle inequality for generalized linear models

First, we remark that the binarized matrix, $\boldsymbol{X}^B$, satisfies $\max_{i=1,\ldots,n} \|\boldsymbol{X}^B_{i,\bullet}\|_2 = \sqrt{p}$. A direct fact of this remark is given in the next Lemma.

Lemma 3.4.5. *One has*

$$\max_{i=1,\ldots,n} \sup_{\theta,\eta \in B_d(R)} |\langle \boldsymbol{X}^B_{i,\bullet}, \theta - \eta \rangle| \leq 2R\sqrt{p}. \tag{3.41}$$

To compare the empirical Kullback divergence and the empirical squared norm, we use Lemma 1 in Bach (2010), that we recall here.

Lemma 3.4.6. *Let φ be a convex three times differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|\varphi'''(t)| \leq M|\varphi''(t)|$, for some $M \geq 0$. Then, for all $t \geq 0$, one has*

$$\frac{\varphi''(0)}{M^2}\psi(-Mt) \leq \varphi(t) - \varphi(0) - \varphi'(0)t \leq \frac{\varphi''(0)}{M^2}\psi(Mt), \text{ with } \psi(u) = e^u - u - 1.$$

Now, we give a version of the previous Lemma for our setting.

Lemma 3.4.7. *There exists a constant $C_n(R,p) = R\sqrt{p} + C_n$, such that*

$$\frac{\xi_n^- \psi(-C_n(R,p))}{C_n(R,p)^2}\frac{1}{n}\|m_0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \leq KL_n(m^0(\boldsymbol{X}), m^\theta(\boldsymbol{X})), \tag{3.42}$$

$$\frac{\xi_n^+ \psi(C_n(R,p))}{C_n(R,p)^2}\frac{1}{n}\|m_0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \geq KL_n(m^0(\boldsymbol{X}), m^\theta(\boldsymbol{X})), \tag{3.43}$$

for all $\theta \in B_d(R)$.

Proof of Lemma 3.4.7

Let us consider the function $G_n : \mathbb{R} \to \mathbb{R}$, defined by $G_n(t) = R_n(m_\vartheta + tm_\eta)$, then

$$G_n(t) = \frac{1}{n}\sum_{i=1}^n b(m_{\vartheta+t\eta}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i m_{\vartheta+t\eta}(\boldsymbol{X}_{i,\bullet}).$$

By differentiating G_n with respect to the variable t we obtain:

$$\dot{G}_n(t) = \frac{1}{n}\sum_{i=1}^n m_\eta(\boldsymbol{X}_{i,\bullet})\dot{b}(m_{\vartheta+t\eta}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i m_\eta(\boldsymbol{X}_{i,\bullet}),$$

$$\ddot{G}_n(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^2(\boldsymbol{X}_{i,\bullet})\ddot{b}(m_{\vartheta+t\eta}(\boldsymbol{X}_{i,\bullet})),$$

$$\dddot{G}_n(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^3(\boldsymbol{X}_{i,\bullet})\dddot{b}(m_{\vartheta+t\eta}(\boldsymbol{X}_{i,\bullet})).$$

Thus

$$|\dddot{G}_n(t)| \leq \|m_\eta\|_\infty |\ddot{G}_n(t)|,$$

where $\|m_\eta\|_\infty := \max_{i=1,\dots,n} |m_\eta(\boldsymbol{X}_{i,\bullet})|$. Applying Lemma 3.4.6 with $M = \|m_\eta\|_\infty$, we obtain for all $t \geq 0$,

$$\ddot{G}_n(0)\frac{\psi(-\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2} \leq G_n(t) - G_n(0) - t\dot{G}_n(0) \leq \ddot{G}_n(0)\frac{\psi(\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2}.$$

Taking $t = 1$, it implies

$$\ddot{G}_n(0)\frac{\psi(-\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2} \leq R_n(m_\vartheta + tm_\eta) - R_n(m_\vartheta) - \dot{G}_n(0),$$

$$\ddot{G}_n(0)\frac{\psi(\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2} \geq R_n(m_\vartheta + tm_\eta) - R_n(m_\vartheta) - \dot{G}_n(0).$$

A short calculation gives that

$$
\begin{aligned}
-\dot{G}_n(0) &= \frac{1}{n}\sum_{i=1}^{n} m_\eta(\boldsymbol{X}_{i,\bullet})\big(\boldsymbol{Y}_i - \dot{b}(m_\vartheta(\boldsymbol{X}_{i,\bullet}))\big) \\
&= \frac{1}{n}\sum_{i=1}^{n} m_\eta(\boldsymbol{X}_{i,\bullet})\big(\boldsymbol{Y}_i - \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\big) + \frac{1}{n}\sum_{i=1}^{n} m_\eta(\boldsymbol{X}_{i,\bullet})\big(\dot{b}(m_0(\boldsymbol{X}_{i,\bullet})) - \dot{b}(m_\vartheta(\boldsymbol{X}_{i,\bullet}))\big), \\
\ddot{G}_n(0) &= \frac{1}{n}\sum_{i=1}^{n} m_\eta^2(\boldsymbol{X}_{i,\bullet})\ddot{b}(m_\vartheta(\boldsymbol{X}_{i,\bullet})).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\ddot{G}_n(0)\frac{\psi(-\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2} &\le \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\vartheta + m_\eta)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\vartheta)] + Q_n(\dot{b}(m_0(\boldsymbol{X})), \dot{b}(m_\vartheta(\boldsymbol{X}))), \\
\ddot{G}_n(0)\frac{\psi(\|m_\eta\|_\infty)}{\|m_\eta\|_\infty^2} &\ge \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\vartheta + m_\eta)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\vartheta)] + Q_n(\dot{b}(m_0(\boldsymbol{X})), \dot{b}(m_\vartheta(\boldsymbol{X}))),
\end{aligned}
$$

where $Q_n(\dot{b}(m_0(\boldsymbol{X})), \dot{b}(m_\vartheta(\boldsymbol{X}))) = \frac{1}{n}\langle \dot{b}(m_0(\boldsymbol{X})) - \dot{b}(m_\vartheta(\boldsymbol{X})), m_\eta(\boldsymbol{X})\rangle$. For $m_\eta = m_\theta - m_0$, and $m_\vartheta = m_0$, we obtain,

$$
\begin{aligned}
\|m_\eta\|_\infty &= \max_{i=1,\dots,n} |\langle \boldsymbol{X}_{i,\bullet}^B, \theta\rangle - m_0(\boldsymbol{X}_{i,\bullet})| \\
&\le \max_{i=1,\dots,n} \big(|\langle \boldsymbol{X}_{i,\bullet}^B, \theta\rangle| + |m_0(\boldsymbol{X}_{i,\bullet})|\big) \\
&\le R\sqrt{p} + C_n \\
&= C_n(R,p).
\end{aligned}
$$

Then, we arrive at

$$
\begin{aligned}
\ddot{G}_n(0)\frac{\psi(-C_n(R,p))}{C(R,p)^2} &\le \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\theta)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_0)] = KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})), \\
\ddot{G}_n(0)\frac{\psi(C_n(R,p))}{C(R,p)^2} &\ge \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\theta)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_0)] = KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),
\end{aligned}
$$

with

$$
\ddot{G}_n(0) = \frac{1}{n}\sum_{i=1}^{n}\big(m_\theta(\boldsymbol{X}_{i,\bullet}) - m_0(\boldsymbol{X}_{i,\bullet})\big)^2\ddot{b}(m_0(\boldsymbol{X}_{i,\bullet})).
$$

It entails that

$$
\begin{aligned}
&\frac{\xi_n^- \psi(-C_n(R,p))}{C_n(R,p)^2}\frac{1}{n}\|m_0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \\
&\qquad \le KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \le \frac{\xi_n^+ \psi(C_n(R,p))}{C_n(R,p)^2}\frac{1}{n}\|m_0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2.
\end{aligned}
$$

□

Proof of Theorem 3.3.7

In this setting, recall that for all $\theta \in \mathbb{R}^d$,

$$
R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} b(m_\theta(\boldsymbol{X}_{i,\bullet}^B)) - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i m_\theta(\boldsymbol{X}_{i,\bullet}^B),
$$

and

$$\hat{\theta} = \operatorname{argmin}_{\theta \in B_d(R)} \{R_n(\theta) + \operatorname{bina}_{\hat{w}}(\theta)\}. \tag{3.44}$$

According to Proposition 3.4.2, Equation (3.44) means that there is $\hat{h} = (\hat{h}_{j,\bullet})_{j=1,\dots,p} \in \partial(\|\hat{\theta}\|_{\mathrm{bTV},\hat{w}})$, $\hat{g} = (\hat{g}_{j,\bullet})_{j=1,\dots,p} \in \big(\partial(\delta_{\mathscr{H}_j}(\hat{\theta}_{j,\bullet}))\big)_{j=1,\dots,p}$, and $\hat{f} = (\hat{f}_{j,\bullet})_{j=1,\dots,p} \in \partial(\delta_{B_d(R)}(\hat{\theta}))$ such that

$$\langle \frac{1}{n}(\boldsymbol{X}^B)^\top \big(\dot{b}(\boldsymbol{X}^B\hat{\theta}) - \boldsymbol{Y}\big) + \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta \rangle = 0, \text{ for all } \theta \in \mathbb{R}^d.$$

Equivalently,

$$\langle \frac{1}{n}(\boldsymbol{X}^B)^\top \big(\dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{Y}\big) + \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta \rangle = 0, \text{ for all } \theta \in \mathbb{R}^d. \tag{3.45}$$

Equation (3.45) can be written in the following way:

$$\frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle - \frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle + \langle \hat{h}, \hat{\theta} - \theta\rangle + \langle \hat{g} + \hat{f}, \hat{\theta} - \theta\rangle = 0. \tag{3.46}$$

Now, using the fact that the subdifferential mapping is monotone (this is an immediate consequence of its definition, see Rockafellar (1970), to say that $\langle \hat{h} - h, \hat{\theta} - \theta\rangle \geq 0$, for any $h \in \partial(\|\theta\|_{\mathrm{bTV},\hat{w}})$ entails that

$$\frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \leq \frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle - \langle h, \hat{\theta} - \theta\rangle - \langle \hat{g} + \hat{f}, \hat{\theta} - \theta\rangle.$$

Furthermore, for any $\theta = (\theta_{j,\bullet}^\top \cdots \theta_{p,\bullet}^\top)^\top \in B_d(R) \cap \mathscr{H}_1 \times \cdots \times \mathscr{H}_p$, we have $-\langle \hat{g}_{j,\bullet} + \hat{f}_{j,\bullet}, \hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\rangle \leq 0$. Hence, for any $\theta \in B_d(R) \cap \mathscr{H}_1 \times \cdots \times \mathscr{H}_p$ and $h \in \partial(\|\theta\|_{\mathrm{bTV},\hat{w}})$, it follows

$$\frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle \leq \frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle - \langle h, \hat{\theta} - \theta\rangle. \tag{3.47}$$

For a fixed $\eta \in B_d(R)$, we consider the function $H_n : \mathbb{R} \to \mathbb{R}$, defined by

$$H_n(t) = \frac{1}{n}\sum_{i=1}^n b(m_{\hat{\theta}+t\eta}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^n \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\langle \boldsymbol{X}_{i,\bullet}^B, \hat{\theta} + t\eta\rangle$$

By differentiating H_n with respect to the variable t we obtain:

$$\begin{aligned}
\dot{H}_n(t) &= \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{X}_{i,\bullet}^B, \eta\rangle \dot{b}(m_{\hat{\theta}+t\eta}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^n \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\langle \boldsymbol{X}_{i,\bullet}^B, \eta\rangle, \\
\ddot{H}_n(t) &= \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{X}_{i,\bullet}^B, \eta\rangle^2 \ddot{b}(m_{\hat{\theta}+t\eta}(\boldsymbol{X}_{i,\bullet})), \\
\dddot{H}_n(t) &= \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{X}_{i,\bullet}^B, \eta\rangle^3 \dddot{b}(m_{\hat{\theta}+t\eta}(\boldsymbol{X}_{i,\bullet})).
\end{aligned}$$

Obviously, we have

$$|\dddot{H}_n(t)| \leq 2R\sqrt{p}|\ddot{H}_n(t)|.$$

Applying Lemma 3.4.6 with $M(R,p) := 2R\sqrt{p}$, we obtain for all $t \geq 0$,

$$\ddot{H}_n(0)\frac{\psi(-M(R,p))}{M(R,p)^2} \leq H_n(t) - H_n(0) - t\dot{H}_n(0) \leq \ddot{H}_n(0)\frac{\psi(M(R,p))}{M(R,p)^2}.$$

Taking $t = 1$, and $\eta = \theta - \hat{\theta}$, implies

$$H_n(1) = \frac{1}{n}\sum_{i=1}^{n} b(m_\theta(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^{n} \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\langle \boldsymbol{X}^B_{i,\bullet}, \theta\rangle = \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\theta)],$$
$$H_n(0) = \frac{1}{n}\sum_{i=1}^{n} b(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^{n} \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\langle \boldsymbol{X}^B_{i,\bullet}, \hat{\theta}\rangle = \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_{\hat{\theta}})].$$

Moreover, we have

$$\begin{aligned}
\dot{H}_n(0) &= \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{X}^B_{i,\bullet}, \theta - \hat{\theta}\rangle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})) - \frac{1}{n}\sum_{i=1}^{n} \dot{b}(m_0(\boldsymbol{X}_{i,\bullet}))\langle \boldsymbol{X}^B_{i,\bullet}, \hat{\theta} - \theta\rangle \\
&= \frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\theta - \hat{\theta})\rangle, \\
\ddot{H}_n(0) &= \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{X}^B_{i,\bullet}, \hat{\theta} - \theta\rangle^2 \ddot{b}(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})).
\end{aligned}$$

Then, we deduce that

$$\begin{aligned}
&\ddot{H}_n(0)\frac{\psi(-M(R,p))}{M(R,p)^2} \\
&\quad \le \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\theta)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_{\hat{\theta}})] - \frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\theta - \hat{\theta})\rangle \\
&\quad \le \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_\theta) - R_n(m_0)] - \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(m_{\hat{\theta}}) - R_n(m_0)] \\
&\qquad\qquad - \frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\theta - \hat{\theta})\rangle \\
&\quad = KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) - KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\
&\qquad\qquad + \frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle.
\end{aligned}$$

Therefore

$$\begin{aligned}
&KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) + \ddot{H}_n(0)\frac{\psi(-M(R,p))}{M(R,p)^2} \\
&\qquad \le KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{1}{n}\langle \dot{b}(m_{\hat{\theta}}(\boldsymbol{X})) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle.
\end{aligned}$$

Then by monotonicity of subdifferential, it implies that

$$\begin{aligned}
&KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) + \ddot{H}_n(0)\frac{\psi(-M(R,p))}{M(R,p)^2} \\
&\qquad \le KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle + \langle h, \hat{\theta} - \theta\rangle.
\end{aligned}$$

Due to the fact that the term $\mathbb{V}[\boldsymbol{Y}|\boldsymbol{X}] = \ddot{b}(m_0(\boldsymbol{X}))$, one has

$$\ddot{H}_n(0) = \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{X}^B_{i,\bullet}, \hat{\theta} - \theta\rangle^2 \ddot{b}(m_{\hat{\theta}}(\boldsymbol{X}_{i,\bullet})) \ge 0.$$

Then

$$\begin{aligned}
&KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\
&\qquad \le KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{1}{n}\langle m_{\hat{\theta}}(\boldsymbol{X}) - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle.
\end{aligned} \tag{3.48}$$

By the monotonicity of the subdifferential, see (3.47), we have that

$$\begin{aligned} &KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &\quad \leq KL_n(m_0(\boldsymbol{X}), m^{\theta}(\boldsymbol{X})) + \frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle - \langle h, \hat{\theta}-\theta\rangle. \end{aligned} \tag{3.49}$$

If $\frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle - \langle h, \hat{\theta}-\theta\rangle < 0$, it follows that $KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X}))$, then the Theorem 3.3.7 holds. From now on, we assume that

$$\frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle - \langle h, \hat{\theta}-\theta\rangle \geq 0. \tag{3.50}$$

Similarly to the Inequality (3.33) in the proof of Theorem 3.3.3, we get

$$\frac{1}{n}\langle \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta}-\theta)\rangle \leq \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{d_j} \left|\langle \left(\boldsymbol{X}^B_{\bullet,j} T_j\right)_{k,\bullet}, \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X}))\rangle\right| \left|\left(D_j\left((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet}\right)\right)_k\right|.$$

Let us define the event

$$\mathscr{Q}_n = \bigcap_{j=1}^{p}\bigcap_{k=2}^{d_j} \mathscr{Q}_{n,j,k}, \text{ where } \mathscr{Q}_{n,j,k} = \left\{\frac{1}{n}\left|\langle \left(\boldsymbol{X}^B_{\bullet,j} T_j\right)_{k,\bullet}, \boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X}))\rangle\right| \leq 2\hat{w}_{j,k}\right\}. \tag{3.51}$$

Then, on $\mathscr{Q}_n$, we have

$$\begin{aligned} \frac{1}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{Y} - \dot{b}(m_0(\boldsymbol{X})), \hat{\theta}-\theta\rangle &\leq 2\sum_{j=1}^{p}\sum_{k=1}^{d_j} \hat{w}_{j,k}\left|\left(D_j\left(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\right)\right)_k\right| \\ &= 2\sum_{j=1}^{p} \|\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \\ &= 2\sum_{j=1}^{p} \|\hat{w}_{j,\bullet} \odot D_j\left(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}\right)\|_1. \end{aligned} \tag{3.52}$$

Anagoulously to the Inequality (3.36) in the proof of Theorem 3.3.3,we get

$$-\langle h, \hat{\theta}-\theta\rangle \leq \sum_{j=1}^{p} \|((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} - \sum_{j=1}^{p} \|((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}}. \tag{3.53}$$

Together Inequalities (3.50), (3.52) and (3.53) with the fact that on $\mathscr{Q}_n$, we have established that

$$\sum_{j=1}^{p} \|((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j^c(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}} \leq 3\sum_{j=1}^{p} \|((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{bTV},\hat{w}_{j,\bullet}},$$

also

$$\sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j^c(\theta)} \odot D_j((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j^c(\theta)}\|_1 \leq 3\sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j((\hat{\theta})_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1. \tag{3.54}$$

This means, (see (3.6) and (3.27)), that

$$\hat{\theta} - \theta \in \mathscr{C}_{\mathrm{bTV},\hat{w}}(J(\theta)), \text{ and } \mathbf{D}(\hat{\theta} - \theta) \in \mathscr{C}_{1,\hat{w}}(J(\theta)).$$

Now, returning to (3.49). Taking into account (3.54), the compatibility of $\boldsymbol{X}^B\mathbf{T}$, see (3.26), and finally the connection between the empirical squared norm and Kullback divergence, see Lemma 3.4.7, on $\mathcal{Q}_n$ we get the following

$$\begin{aligned} KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) &\leq KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + 3\sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1 \\ &\leq KL_n(m^0(\boldsymbol{X}), m^{\theta}(\boldsymbol{X})) + \frac{\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2}{\sqrt{n}\,\kappa_{\mathbf{T},\delta}(J(\theta))\kappa_{\boldsymbol{X}^B}(3, J(\theta))}, \end{aligned}$$

where $\hat{\gamma} = (\hat{\gamma}_{j,\bullet})_{j=1,\dots,p}$ such that

$$\hat{\gamma}_{j,k} = \begin{cases} 3\hat{w}_{j,k}, & \text{if } k \in J_j(\theta), \\ 0, & \text{if } k \in J_j^c(\theta). \end{cases}$$

Thus we have

$$\begin{aligned} &\frac{\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2}{\sqrt{n}\,\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))} \\ &\qquad \leq \frac{1}{\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))}\Big(\frac{1}{\sqrt{n}}\|m_{\hat{\theta}}(\boldsymbol{X}) - m_0(\boldsymbol{X})\|_2 + \frac{1}{\sqrt{n}}\|m_0(\boldsymbol{X}) - m_{\hat{\theta}}(\boldsymbol{X})\|_2\Big). \end{aligned}$$

Using the connection between the empirical norm and the Kullback divergence, see Lemma 3.4.7 we have

$$\begin{aligned} &\frac{\|\boldsymbol{X}^B(\hat{\theta}-\theta)\|_2}{\sqrt{n}\,\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))} \\ &\leq \frac{2C_n(R,p)}{\sqrt{\xi_n^- \psi(-C_n(R,p))}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))}\Big(\sqrt{KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X}))} \\ &\qquad\qquad + \sqrt{KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X}))}\Big) \\ &\leq \frac{2}{\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa_{\boldsymbol{X}^B}(J(\theta))}\Big(\sqrt{C_n(R,p,\xi)^{-1} KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X}))} \\ &\qquad\qquad + \sqrt{C_n(R,p,\xi)^{-1} KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X}))}\Big), \end{aligned}$$

where $C_n(R,p,\xi) = \xi_n^- \psi(-C_n(R,p))/C_n^2(R,p)$. We now use the elementary inequality $2uv \leq \epsilon u^2 + v^2/\epsilon$ with $\epsilon > 0$. Therefore, we have

$$\begin{aligned} &KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &\leq KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + \frac{2\epsilon}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2_{\boldsymbol{X}^B}(J(\theta))} \\ &\qquad + \big(\epsilon C_n(R,p,\xi)\big)^{-1} KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) + \big(\epsilon C_n(R,p,\xi)\big)^{-1} KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})). \end{aligned}$$

By choosing $\epsilon > 1/C_n(R,p,\xi)$, we get

$$\begin{aligned} &KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &\leq \frac{1 + \big(\epsilon C_n(R,p,\xi)\big)^{-1}}{1 - \big(\epsilon C_1(R,p,\xi_0)\big)^{-1}} KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + \frac{2\epsilon^2 C_n(R,p,\xi)}{\big(\epsilon C_n(R,p,\xi) - 1\big)\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2_{\boldsymbol{X}^B}(J(\theta))} \\ &\leq \frac{\epsilon C_n(R,p,\xi) + 1}{\epsilon C_n(R,p,\xi) - 1} KL_n(m_0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + \frac{2\epsilon^2 C_n(R,p,\xi)}{(\epsilon C_n(R,p,\xi) - 1)\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2_{\boldsymbol{X}^B}(J(\theta))}. \end{aligned}$$

We have

$$\frac{1}{\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))} \leq 608|J(\theta)| \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.$$

Setting

$$\frac{\epsilon C_n(R,p,\xi)+1}{\epsilon C_n(R,p,\xi)-1} = 1 + \frac{2}{\epsilon C_n(R,p,\xi)-1} =: 1+\gamma.$$

Then

$$KL_n(m_0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq (1+\gamma) KL_n(m_0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{1216\,\epsilon^2 C_n(R,p,\xi)}{\big(\epsilon C_n(R,p,\xi)-1\big)\kappa^2_{\boldsymbol{X}^B}(J(\theta))} |J(\theta)| \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty.$$

Finally, using an elementary bound on the tails of subgaussian random variables given by Hoeffding inequality, and the choice of the weights $\hat{w}_{j,k}$ for all $j = 1,\dots,p$, and $k = 2,\dots,d_j$, given by (3.12), we find that the probability of the complementary event $\mathcal{Q}_n^c$ is equal to p^{1-A}. This achieves the proof.

□

Chapter 4

Time-Varying High-Dimensional Aalen and Cox Models

This chapter is a preprint of Alaya et al. (2016a).

Abstract

In high dimensional covariates setting, we consider the problem of estimating the intensity of a counting process in the time-varying Aalen and Cox models. We introduce a covariate-specific weighted total variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. We prove theoretical guaranties and we present a proximal algorithm to solve the convex studied problems. The practical use and effectiveness of the proposed method are demonstrated by simulation studies and a real data example.

Contents

4.1 Introduction

4.1.1 Framework and models

Consider the usual counting process framework where a process $\tilde{N}$ counts the number of occurring events of interest over a fixed time interval, say $[0,\tau]$ with $0<\tau<\infty$, and the convention $\tilde{N}(0)=0$ (see Andersen et al. (1993); Martinussen and Scheike (2007)) . Let $\lambda_\star$ denote the intensity of the process $\tilde{N}$ depending on both time and a p-dimensional predictable process of covariates denoted by X (possibly including an intercept).

We consider that the process $\tilde{N}$ may be independently filtered (see Andersen et al. (1993)) by a censoring predictable process Y and the resulting observed process is denoted by N. The intensity of N is then given for all $t\in[0,\tau]$ by

$$Y(t)\lambda_\star(t,X(t)).$$

In this framework, we consider two dynamic models for the function $\lambda_\star$:

— a time-varying Aalen model

$$\lambda_\star^{\mathrm{A}}(t,X(t))=X(t)\beta^\star(t), \tag{4.1}$$

— a time-varying Cox model

$$\lambda_\star^{\mathrm{M}}(t,X(t))=\exp\left(X(t)\beta^\star(t)\right) \tag{4.2}$$

where, in both cases, $\beta^\star$ is an unknown p-dimensional function from $[0,\tau]$ to $\mathbb{R}^p$ to be estimated. We consider the problem of estimating the parameter $\beta^\star$ in dynamic models (4.1) and (4.2) on the basis of data from n independent individuals:

$$\mathscr{D}_n=\left\{(X_i(t),Y_i(t),N_i(t)):t\in[0,\tau],i=1,\dots,n\right\}. \tag{4.3}$$

Estimation in models (4.1) and (4.2) are received a lot of attention in the past four decades. References for the additive Aalen model include Aalen (1980, 1989, 1993); Huffer and McKeague (1991); McKeague (1988). For the time-varying Cox models include Cai and Sun (2003); Grambsch and Therneau (1994); Martinussen et al. (2002); Murphy and Sen (1991); Winnett and Sasieni (2003); Zucker and Karr (1990) and very recently Honda and Härdle (2014). In Martinussen and Scheike (2007) may be found a complete presentation of the models, estimation methods and results. The R package `timereg`, see Appendix C in Martinussen and Scheike (2007), implements these procedures.

These models are extensions of the classical Aalen (1980) and Cox (1972) models with constant regression parameters. Dynamic models are obviously more flexible than their constant counterparts, but they suffer from their complexities: p unknown functions are to be estimated from the data. We propose in the present paper a penalized procedure that will reaches a comprise between these two extreme situations, and in addition perform variable selection.

4.1.2 Penalized piecewise constant estimators

Following Murphy and Sen (1991), we consider sieves (or histogram) based estimators of the p-dimensional unknown function $\beta^\star$. We hence consider a L-partition of the time interval $[0,\tau]$, where $L \in \mathbb{N}^*$,

$$\varphi_l = \sqrt{\frac{L}{\tau}} \mathbb{1}(I_l) \text{ and } I_l = \Big(\frac{l-1}{L}\tau, \frac{l}{L}\tau\Big]. \tag{4.4}$$

For all $j = 1,\dots,p$, candidates for the estimation of the j-th coordinate $\beta_j^\star$ of $\beta^\star$ belongs to the set of univariate piecewise constant functions

$$\mathscr{H}_L = \Big\{\alpha(\cdot) = \sum_{l=1}^{L} \alpha_l \varphi_l(\cdot) : (\alpha_l)_{1\le l\le L} \in \mathbb{R}_+^L\Big\}. \tag{4.5}$$

For moderate sample size n and/or high dimensional covariates and/or a fine partition, the resulting estimators would suffer from over-parametrization, in the sense that $\sqrt{n}$ could be much lesser than $p \times L$. On the other hand, simpler forms of models (4.1) and (4.2), when the functions $\beta_j^\star$ are constant over $[0,\tau]$, are often to poor to accurately fit the data (see the discussions on page 205 and following in Martinussen and Scheike (2007) and in Paragraph 4.1.1).

We here seek to reach a compromise between these two extreme situations by introducing a covariate specific weighted $\ell_1+\ell_1$-total-variation penalty (defined in (4.9)). The total-variation part in the penalty induces simple, interpretable estimators, which do not vary much over the time. The ℓ_1 part allows our procedure to support high-dimensional (with a large p) covariates.

Our algorithms are part of the class of fused Lasso algorithms. The latter have been introduced and studied, for noised piecewise constant signals, by Tibshirani et al. (2005), Rinaldo (2009), Harchaoui and Lévy-Leduc (2010), or Dalalyan et al. (2014). A total-variation penalized estimator has been investigated in Alaya et al. (2015) for estimating the intensity of a counting process, while Alaya et al. (2016b); Bouaziz and Guilloux (2015) proposed related estimators in other contexts.

Lasso estimators in the context of survival analysis with high dimensional covariates have been introduced and studied in Gaïffas and Guilloux (2012); Martinussen and Scheike (2009) in the Aalen model and Huang et al. (2013); Lemler (2013); Tibshirani (1997) in the Cox model, among others.

The main contributions of the paper are the following.

— Theoretical: We propose new estimators, see the definitions in (4.12) and (4.11) for the problem at hand. We investigate their theoretical properties by proving oracle inequalities, staten in Section 4.3, that assure their convergences. We also exhibit a theoretical order of magnitude for the choice of the partition's size L, see Section 4.8 in the supplementary material, which is used in the implementation.

— Practical: We propose new algorithm, see Section 4.4, for computing our estimators in the dynamic models of Equations (4.1) and (4.2). We demonstrate in Section 4.8 that they outperform existing algorithms both in terms of estimation precision and complexity. More importantly, our algorithms are of the

class of stochastic gradient descent ones (see Bottou (2010); Bousquet and Bottou (2008)) and are, as such, scalable.

The paper is organized as follows. Section 4.1.2 is devoted to the definition of our estimators and the statement of their theoretical properties. In Section 4.4, we describe our algorithms. Simulation results and illustration on real data are presented in Section 4.5. Section 4.6.2 is a supplementary materials devoted to prove other theoretical results.

4.2 Estimation procedures

We describe in this section our novel estimation procedures, which involve a $\ell_1 + \ell_1$-total-variation penalization of criteria, specific to either the multiplicative or additive models.

4.2.1 Estimation

Estimation in traditional constant coefficients models is based on a minimization of a (partial) least-square criterion in the usual Aalen (1980) model and a (partial) log-likelihood maximization in the usual Cox (1972) model. We refer the reader to Gaïffas and Guilloux (2012); Martinussen and Scheike (2009) and Cox (1975) for the details.

We now introduce some notations. For each individual i with a p-dimensional process of covariates X_i, we denote by X_i^j the process associated to its j-th covariate. Accordingly, for any p-dimensional function β, candidate for the estimation of $\beta^\star$, the univariate function β_j is its j-th coordinate. We define the sets of candidates for estimation as

$$\Lambda^{\mathrm{A}} = \{x, t \in [0,\tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = x(t)\beta(t) \mid \forall j\ \beta_j \in \mathscr{H}_L\}. \tag{4.6}$$

for the Aalen model (4.1) and

$$\Lambda^{\mathrm{M}} = \{x, t \in [0,\tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = \exp\big(x(t)\beta(t)\big) \mid \forall j\ \beta_j \in \mathscr{H}_L\}. \tag{4.7}$$

for the Cox model (4.2).

As for a candidate in Λ^{A} or Λ^{M}, each time-varying coefficient β_j is piecewise constant, we will refer equivalently to β as a p-dimensional function or as the vector of dimension $p \times L$

$$\beta = (\beta_{1,\cdot}^\top, \ldots, \beta_{p,\cdot}^\top)^\top = (\beta_{1,1}, \ldots, \beta_{1,L}, \ldots, \beta_{p,1}, \ldots, \beta_{p,L})^\top,$$

where $\beta_{j,\cdot}$ is in $\mathbb{R}^L$ and $\beta_{j,l}$ is the value taken by the j-th coordinate on the l-th time interval in our L-partition $\{I_1, \ldots, I_L\}$:

$$\forall j = 1\ldots, p,\ \forall l = 1, \ldots, L \text{ and } \forall t \in I_l, \beta_j(t) = \sqrt{\frac{L}{\tau}}\beta_{j,l}.$$

Estimation in the time varying Cox and Aalen models

The existing estimators in the dynamic additive (4.1) and multiplicative (4.2) models with time varying coefficients are also defined via respectively the least-squares

(see page 108 and following in Martinussen and Scheike (2007))) and log-likelihood (see page 206 and following in Martinussen and Scheike (2007)) criteria.

The time-varying Aalen model

For the time-varying Aalen model, we consider the least square criterion for our data and a candidate $\lambda_\beta^{\mathrm{A}}$ defined by

$$\ell_n^{\mathrm{A}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \big(\lambda_\beta^{\mathrm{A}}(t,X_i(t))\big)^2 Y_i(t)dt - 2\int_0^\tau \lambda_\beta^{\mathrm{A}}(t,X_i(t))dN_i(t)\Big\}, \tag{4.8}$$

see Gaïffas and Guilloux (2012) for details on this criterion. When the candidate $\lambda_\beta^{\mathrm{A}}$ is in the class Λ^{A}, Equation (4.8) simplifies to

$$\begin{aligned}\ell_n^{\mathrm{A}}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \big(\sum_{j=1}^{p} X_i^j(t)\beta_j(t)\big)^2 Y_i(t)dt - 2\int_0^\tau \sum_{j=1}^{p} X_i^j(t)\beta_j(t)dN_i(t)\Big\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\Big\{\frac{L}{\tau}\int_{I_l}\big(\sum_{j=1}^{p} X_i^j(t)\beta_{j,l}\big)^2 Y_i(t)dt - 2\sqrt{\frac{L}{\tau}}\sum_{j=1}^{p}\big(\int_{I_l} X_i^j(t)dN_i(t)\big)\beta_{j,l}\Big\}.\end{aligned}$$

The time varying Cox model

We consider, in this paragraph, estimation in the time-varying Cox model. Minus the log-likelihood for our data and a candidate $\lambda_\beta^{\mathrm{M}}$ is given by

$$\ell_n^{\mathrm{M}}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\big(\lambda_\beta^{\mathrm{M}}(t,X_i(t))\big)dN_i(t) - \int_0^\tau Y_i(t)\lambda_\beta^{\mathrm{M}}(t,X_i(t))dt\Big\},$$

see Andersen et al. (1993) for details. When $\lambda_\beta^{\mathrm{M}}$ is in the class Λ^{M}, the last expression reduces to

$$\begin{aligned}\ell_n^{\mathrm{M}}(\beta) &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \sum_{j=1}^{p} X_i^j(t)\beta_j(t)dN_i(t) - \int_0^\tau Y_i(t)\exp\big(\sum_{j=1}^{p} X_i^j(t)\beta_j(t)\big)dt\Big\}\\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\Big\{\sqrt{\frac{L}{\tau}}\sum_{j=1}^{p}\Big(\int_{I_l} X_i^j(t)dN_i(t)\Big)\beta_{j,l} - \int_{I_l} Y_i(t)\exp\Big(\sqrt{\frac{L}{\tau}}\sum_{j=1}^{p} X_i^j(t)\beta_{j,l}\Big)dt\Big\}.\end{aligned}$$

Estimation procedure

We introduce a well-chosen vector of data-driven weights $\hat{\gamma} = (\hat{\gamma}_{1,\cdot}^\top, \ldots, \hat{\gamma}_{p,\cdot}^\top)^\top$ of order (we write here only the dominating terms, see Definition 4.6.1 in supplementary material for its explicit form)

$$\hat{\gamma}_{j,l} \asymp \sqrt{\frac{x + L\log pL}{n}\hat{V}_{j,l}},$$

where

$$\hat{V}_{j,l} = \frac{1}{n}\sum_{i=1}^{n}\int_{(l-1)\tau/L}^{\tau}(X_i^j(t))^2 dN_i(t),$$

and our covariate specific weighted $\ell_1 + \ell_1$-total-variation penalty

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \sum_{j=1}^{p} \left(\hat{\gamma}_{j,1} |\beta_{j,1}| + \sum_{l=2}^{L} \hat{\gamma}_{j,l} |\beta_{j,l} - \beta_{j,l-1}| \right) \tag{4.9}$$

for $\beta \in \mathbb{R}^{p \times L}$. Our estimators are then respectively defined as

$$\hat{\lambda}^{\mathrm{A}} = \lambda^{\mathrm{A}}_{\hat{\beta}^{\mathrm{A}}}, \text{ and } \hat{\lambda}^{\mathrm{M}} = \lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}. \tag{4.10}$$

where

$$\hat{\beta}^{\mathrm{A}} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p \times L}} \left\{ \ell^{\mathrm{A}}_n(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right\}, \tag{4.11}$$

in the Aalen model and

$$\hat{\beta}^{\mathrm{M}} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p \times L}} \left\{ \ell^{\mathrm{M}}_n(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right\}, \tag{4.12}$$

in the Cox model.

4.3 Theoretical guaranties

In this section we address the statistical properties of the weighted $\ell_1 + \ell_1$-total-variation estimation procedure presented in the previous section. Our first results establish theoretical properties of our estimators by using the classical non-asymptotic oracle approaches. Towards this end, we first introduce the weighted empirical quadratic norm $\|\lambda^{\mathrm{A}}\|_n$ defined for any $\lambda^{\mathrm{A}} \in \Lambda^{\mathrm{A}}$ by

$$\|\lambda^{\mathrm{A}}\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\lambda^{\mathrm{A}}(t, X_i(t))\right)^2 Y_i(t) dt},$$

and the empirical Kullback divergence $K_n(\lambda^{\mathrm{M}}_\star, \lambda^{\mathrm{M}}_\beta)$ defined for $\lambda^{\mathrm{M}}_\beta \in \Lambda^{\mathrm{M}}$ by

$$\begin{aligned} K_n(\lambda^{\mathrm{M}}_\star, \lambda^{\mathrm{M}}_\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\log \lambda^{\mathrm{M}}_\star(t, X_i(t)) - \log \lambda^{\mathrm{M}}_\beta(t, X_i(t)) \right) \lambda^{\mathrm{M}}_\star(t, X_i(t)) Y_i(t) dt \\ - \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\lambda^{\mathrm{M}}_\star(t, X_i(t)) - \lambda^{\mathrm{M}}_\beta(t, X_i(t)) \right) Y_i(t) dt. \end{aligned}$$

Theorem 4.3.1. *For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{A}}$ defined in (4.11), verifies with a probability exceeding $1 - C_A e^{-x}$, for a some constant $C_A > 0$,*

$$\|\lambda^{\mathrm{A}}_\star - \hat{\lambda}^{\mathrm{A}}\|^2_n \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \left(\|\lambda^{\mathrm{A}}_\star - \lambda^{\mathrm{A}}_\beta\|^2_n + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right). \tag{4.13}$$

Theorem 4.3.2. *For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{M}}$ defined in (4.12), verifies with a probability larger than $1 - C_M e^{-x}$, for a constant $C_M > 0$,*

$$K_n(\lambda^{\mathrm{M}}_\star, \hat{\lambda}^{\mathrm{M}}) \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \left(K_n(\lambda^{\mathrm{M}}_\star, \lambda^{\mathrm{M}}_\beta) + 2||\beta||_{\mathrm{gTV},\hat{\gamma}} \right). \tag{4.14}$$

The proofs of Theorems 4.3.1 and 4.3.2 are presented in Section 4.6.1 and 4.6.2 respectively. Two terms are involved on the right hand side of (4.13) and (4.14). The first one measures how far are the true functions of interest $\lambda_\star^{\mathrm{A}}$ and $\lambda_\star^{\mathrm{M}}$ from their approximations on Λ^{A} and Λ^{M}. The second one can be viewed as a variance term that satisfies

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \asymp \|\beta\|_{\mathrm{gTV}} \max_{j=1,\dots,p} \max_{l=1,\dots,L} \sqrt{\frac{L\log pL}{n}\hat{V}_{j,\ell}}. \tag{4.15}$$

for any $\beta \in \mathbb{R}^{p\times L}$. Here, $\|\cdot\|_{\mathrm{gTV}}$ stands for the unweighted $\ell_1+\ell_1$-total variation ($\hat{\gamma}_{j,l}=1$). The dominant term in (4.15) is, up to the loglog term, of order $\|\beta\|_{\mathrm{gTV}}\big(L\log(pL)/n\big)^{1/2}$, which is the expected slow rate for $\hat{\lambda}^{\mathrm{A}}$ and $\hat{\lambda}^{\mathrm{M}}$ involving the total-variation penalization. Such oracle inequality is now classical in the huge literature of the sparsity procedures see for instance Alaya et al. (2015); Bickel et al. (2009); Bunea et al. (2007); Gaïffas and Guilloux (2012); Hansen et al. (2015); van de Geer and Bühlmann (2009). Most of these papers aim at establishing oracle inequalities under weak assumptions on the design matrix.

Theorems 4.3.1 and 4.3.2 are slow oracle inequalities, which are obtained without any assumption. To establish a fast oracle inequalities, the *restricted eigenvalue* (RE) assumption is required. The RE excludes strong correlations between covariables and it was introduced in Bickel et al. (2009). In van de Geer and Bühlmann (2009), one can find an exhaustive survey and comparison of the assumptions used to prove fast oracle inequalities.

In supplementary material, Theorems 4.7.3 and 4.7.7 are fast oracle inequalities for $\lambda_\star^{\mathrm{A}}$ and $\lambda_\star^{\mathrm{M}}$ respectively. Furthermore, under the assumption that the true regression function $\beta^\star$ is piecewise constant, we have an optimal trade-off between approximation and complexity given by the choice $L = O(n^{1/3})$, (for more details see Section 4.8 in the supplementary materials).

4.4 Algorithm

In this section, we use stochastic proximal gradient descent (SPGD) (see Mairal (2013)) for computing solutions of the regularized problems (4.10) and (4.12). We begin by reviewing SPGD in generality, then we describe its implementation for our problem.

We are interested in computing a solution

$$x^\star = \operatorname{argmin}_{x\in\mathbb{R}^p}\{g(x)+h(x)\}, \tag{4.16}$$

where g is the average of the smooth convex functions $g_1,\dots,g_n$ from $\mathbb{R}^p$ to $\mathbb{R}$, i.e., $g(x)=\frac{1}{n}\sum_{i=1}^n g_i(x)$ and $h:\mathbb{R}^p\to\mathbb{R}$ is a relative simple convex function that can be non-differentiable. To solve the optimization problem (4.16), one popular method is proximal gradient descent method (PGD) which can be described by the following update rule for $k=1,2,\dots$:

$$x^{(k+1)} = \operatorname{prox}_{\epsilon_k h}\left(x^{(k)} - \epsilon_k \nabla g(x^{(k)})\right). \tag{4.17}$$

A stochastic variant of PGD is SPGD , where at each iteration $k = 1, 2, \ldots$, we pick i_k randomly from $\{1, 2, \ldots, n\}$, and take the following update:

$$x^{(k+1)} = \text{prox}_{\epsilon_k h}\left(x^{(k)} - \epsilon_k \nabla g_{i_k}(x^{(k)})\right)$$

PGD uses all of the training instances, by evaluation n gradients, to update the model parameters in each iteration. In contrast, SPGD updates the parameters using only a single training instance in each iteration, $\nabla g_{i_k}(x^{(k)})$. The training instance is usually selected randomly. Thus the computational cost of SPGD per iteration is $1/n$ that of the PGD. SPGD is often preferred to optimize cost functions when there are hundreds of thousands of training instances or more, as it will converge more quickly than PGD.

4.4.1 Applications to Aalen and Cox time-varying models

In Algorithm 6, we need the proximal operator of the weighted $\ell_1 + \ell_1$-total variation, namely

$$\theta = \text{argmin}_{x \in \mathbb{R}^{p \times L}} \frac{1}{2}\|\beta - x\|_2^2 + \sum_{j=1}^{p}\left(\gamma_{j,1}x_{j,1} + \sum_{l=2}^{L}\gamma_{j,l}|x_{j,l} - x_{j,l-1}|\right).$$

For more details on this proximal operator, we refer to Alaya et al. (2016b). Since $\|\cdot\|_{\text{gTV},\gamma}$ is separable by blocks, we have

$$\left(\text{prox}_{\|\cdot\|_{\text{gTV},\gamma}}(\beta)\right)_{j,\cdot} = \text{prox}_{\|\cdot\|_{\text{gTV},\gamma}}(\beta_{j,\cdot})$$

for all $j = 1, \ldots, p$. Thus, let us focus on a single j-th block. Algorithm 5 expresses $\text{prox}_{\|\cdot\|_{\text{gTV},\gamma}}(\beta)$ basing on the proximal operator of the weighted total-variation penalization, namely $\text{prox}_{\|\cdot\|_{\text{TV},w}}$. We refer to Alaya et al. (2015) where the authors gave an algorithm for $\text{prox}_{\|\cdot\|_{\text{TV},w}}$. In Algorithm 6, we implement PSGD algorithm via vSGD procedure, see Schaul et al. (2012). Note that in the pseudo-code below, the odot $\odot$ product and division are element-wise operations.

Algorithm 5: $\theta = \text{prox}_{\|\cdot\|_{\text{gTV},\gamma}}(\beta)$

for $j = 1, \ldots, p$ **do**
  set $\mu \leftarrow \beta_{j,\cdot}$; $w \leftarrow \gamma_{j,\cdot} \setminus \{\gamma_{j,1}\}$;
  $\eta \leftarrow \text{prox}_{\|\cdot\|_{\text{TV},w}}(\mu)$;
  $\theta_{j,\cdot} \leftarrow \eta - \left(\eta_1 - \text{sign}(\eta_1)\max\left(0, |\eta_1| - \frac{w_1}{L}\right)\right)\mathbf{1}_L$;
  return $\theta_{j,\cdot}$

For the update of the learning rate ϵ_k in Algorithm 6, we can use the decreasing one of Bottou (2012), i.e. $\epsilon_k = \epsilon_0 \frac{1}{(1+\epsilon_0 \xi k)^\alpha}$ where $\alpha \in [\frac{1}{2}, 1], \xi > 0$, and ϵ_0 learned on small sample of data. Different approaches propose an adaptive update of ϵ_k, for instance AdaGrad in Duchi et al. (2011), AdaDelta in Zeiler (2012) or vSGD in Schaul et al. (2012). In our case, we implement the last one with an explicit computation of the Hessian diagonal of the log-likelihood (others methods use a finite difference approximation of it).

Algorithm 6: SPGD for time-varying Aalen and Cox models

1. Parameters: Integer $K > 0$;
2. Initialization: $(\hat{\beta})^{(1)} = 0 \in \mathbb{R}^{p\times L}$, and $r^{(1)} \in [0,1]$;
3. **for** $k = 1,\dots,K$ **do**
   - Choose randomly $i_k \in \{1,\dots,n\}$ and compute $\nabla^{\mathrm{M}}_{i_k} = \nabla \ell_{i_k}((\hat{\beta})^{(k)})$;
   - Update moving averages
   - $a^{(k)} \leftarrow (1-(r^{(k)})^{-1})a^{(k)} + (r^{(k)})^{-1}\nabla_{i_k}$;
   - $b^{(k)} \leftarrow (1-(r^{(k)})^{-1})b^{(k)} + (r^{(k)})^{-1}\big(\nabla_{i_k}\big)^2$;
   - $c^{(k)} \leftarrow (1-(r^{(k)})^{-1})c^{(k)} + (r^{(k)})^{-1}\operatorname{diag}(H_{i_k})$ where $H_{i_k} = \Big(\frac{\partial^2\left(\ell_{i_k}((\hat{\beta})^{(k)})\right)}{\partial^2\beta}\Big)$;
   - Estimate learning rate
   - $\epsilon^{(k)} \leftarrow \frac{a^{(k)}\odot a^{(k)}}{b^{(k)}\odot c^{(k)}} \in \mathbb{R}^{p\times L}$;
   - **for** $j = 1,\dots,p$ **do**
     - $\eta_j \leftarrow \min_{1\le l\le L} \epsilon^{(k)}_{j,\cdot}$;
   - $\epsilon^{(k)} \leftarrow \big(\eta_1 \mathbf{1}_L,\dots,\eta_p\mathbf{1}_L\big)^\top$;
   - Update memory size
   - $r^{(k)} \leftarrow \big(1 - \frac{a^{(k)}\odot a^{(k)}}{b^{(k)}}\big)\odot r^{(k)} + 1$;
   - $(\hat{\theta})^{(k)} \leftarrow (\hat{\beta})^{(k)} - \epsilon^{(k)}\odot\nabla_{i_k}$;
   - $(\hat{\beta})^{(k+1)} \leftarrow \Big(\operatorname{prox}_{\eta_1\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}}\big((\hat{\theta})^{(k)}_{1,\cdot}\big),\dots,\operatorname{prox}_{\eta_p\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}}\big((\hat{\theta})^{(k)}_{p,\cdot}\big)\Big)^\top$;
4. **return** $(\hat{\beta})^{(K)}$

4.5 Numerical experiments

4.5.1 Simulated data in the time-varying Cox model

We conduct simulations in the time-varying Cox model for survival time with sample size $n = 1000$. The time of interest T has hazard rate $\lambda^\star(t,X) = \beta_0^\star(t)\exp(X(t)\beta^\star(t))$. The $p = 10$ covariates processes $X_i(t)_{i=1,\dots,n}$ are piecewise constant over a 50-partition (as defined in (4.4)) of the time interval $[0,3]$, and each $X_i(t)$ are independent and identically distributed Gaussian random variables, $\mathcal{N}(0,0.5)$. The baseline $\beta_0^\star$ is defined through a Weibull distribution with shape parameter $a = 1.2$ and a scale parameter $b = 0.15$. In Figure 4.1, we draw the true regression functions $\beta_1^\star, \beta_2^\star$, and $\beta_3^\star$. We set $\beta_j^\star \equiv 0$, for $j = 4,\dots,10$.

Given $t \mapsto \lambda^\star(t, X_i(t))$ and $t \mapsto X_i(t)$, the times T_i are simulated as the first event of a non-homogeneous Poisson process with intensity $\lambda^\star(t,X_i(t))$ using thinning (see Lewis and Shedler). Furthermore, we perform a 3-fold cross-validation to select the best constant to use in front of the weights $\hat{\gamma}_{j,\cdot}$.

To evaluate the performance of the estimator, we run 100 Monte-Carlo experiments of training data as described above. The estimation accuracy is investigated via a mean squared error defined as

$$\mathrm{MSE}_j = \frac{1}{100}\sum_{m=1}^{100} \|(\hat{\beta}^{\mathrm{M}}_j)_m - \beta_j^\star\|_2^2,$$

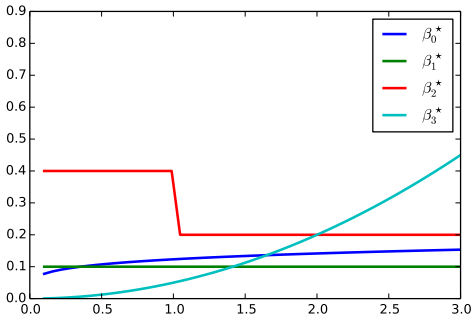


Fig. 4.1 – Baseline and true regression coefficients.

where $(\hat{\beta}_j^{\mathrm{M}})_m$ is the estimation of $\beta_j^\star$ in the sample m, for all $j = 1,\ldots,p,$. In Figure 4.2, we plot the MSE_j of estimated regression coefficients with the $\ell_1 + \ell_1$-weighted total variation proposed procedure (called CoxSGD in the sequel) and the `timereg` R-package (Martinussen and Scheike (2007) over L-partition, where $L \in \{10,30,50,70\}$). In this figure, we observe that CoxSGD performs better than the `timereg` R-package in terms of mean prediction errors. We remark also that for $j = 4,\ldots,p$, the estimators given by CoxSGD do not deviate significantly from zero which is not the case for `timereg` R-package estimators. Our estimation hence performs automatic variable selection.

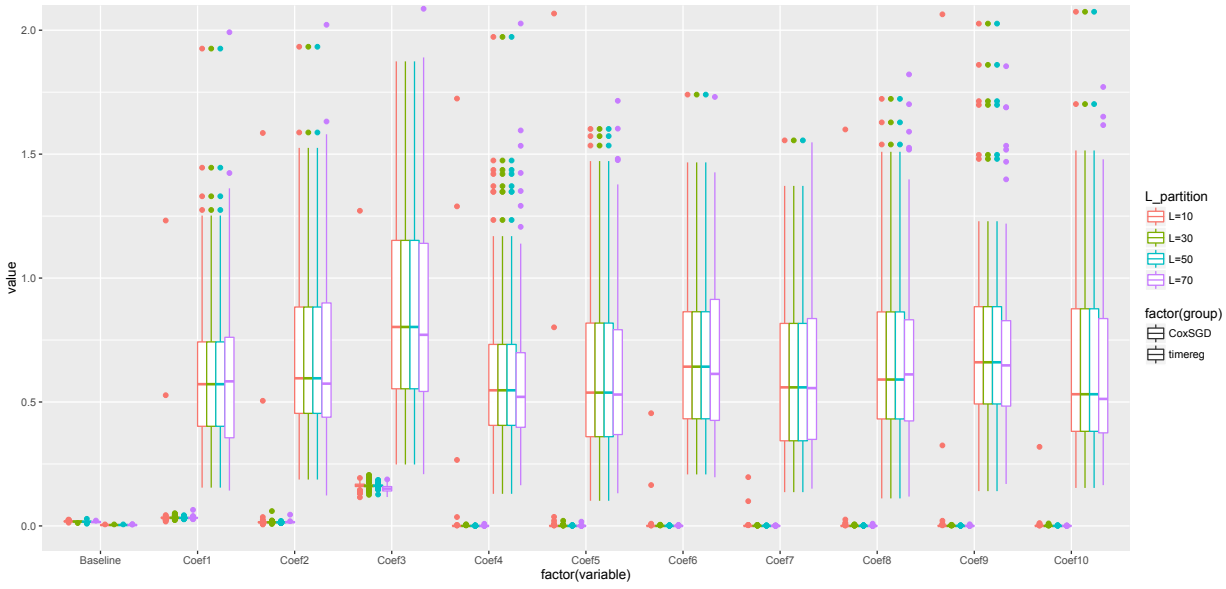


Fig. 4.2 – A boxplots of the MSE_j of estimated regression coefficients over L-partition ($L \in \{10,30,50,70\}$) with CoxSGD and `timereg` R-package.

4.5.2 Real data: illustration using the time-varying Cox model

Our method is illustrated on PBC dataset described in Fleming and Harrington (1991), and originates from a Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver and was conducted between 1974 and 1984. A total of 418 patients are included in the dataset and were followed until death or censoring. We restrict attention to the first 8 years days of the study, and we consider the covariates: age, edema, log(bilirubin), log(albumin) and log(protime). All covariates are centered around their averages. In Figure 4.3, we fit the estimated cumulative regression coefficients ($\hat{B}_j^{\mathrm{M}}(t) = \int_0^t \hat{\beta}_j(s)ds$) on PBC data using CoxSGD and `timereg` R-package. We observe in this figure that $\ell_1 + \ell_1$-total variation procedure gives more interpretable estimators than `timereg` R-package. An important fact is that the runtime of CoxSGD algorithm is extremely fast. Finally, some developments of the present work would be to compare the performance of CoxSGD with fast iterative shrinkage-thresholding procedure (see Beck and Teboulle (2009a)), and the estimator proposed in Winnett and Sasieni (2003).

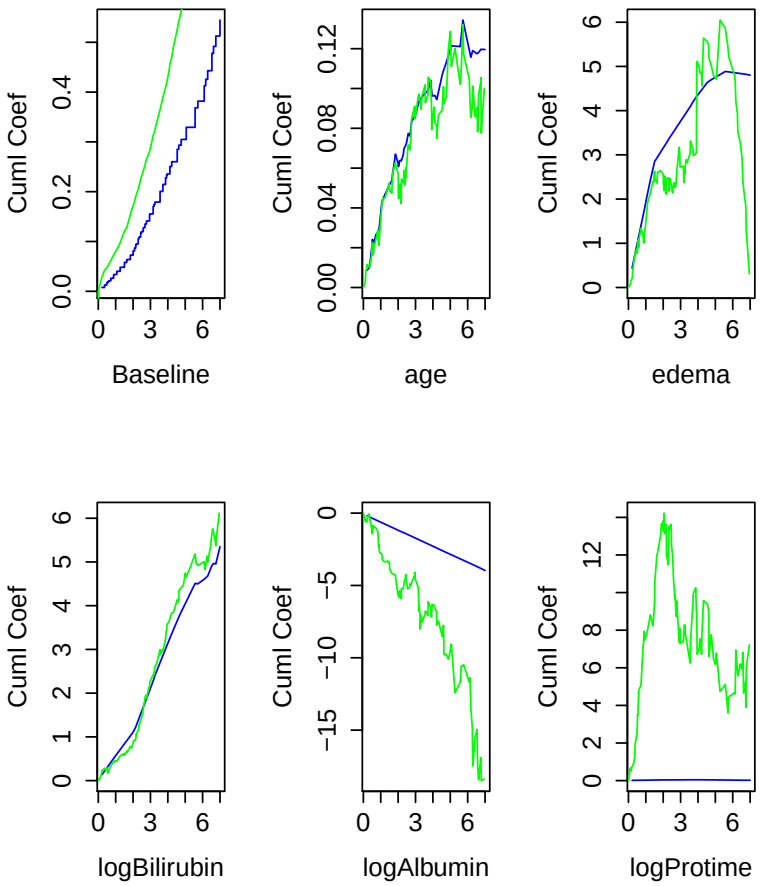


Fig. 4.3 – Estimated cumulative regression coefficients on PBC data: with CoxSGD (blue) and `timereg` R-package (green).

4.6 Proofs

This section is devoted to the proofs of the main results in the paper.

4.6.1 Proof of Theorem 4.3.1: slow oracle inequality in the time-varying Aalen model

Note that for all $t \in [0,\tau]$, $\lambda_\beta^{\mathrm{A}}(t) = \boldsymbol{X}(t)\beta$ where $\boldsymbol{X}(t) = (\boldsymbol{X}_1(t),\dots,\boldsymbol{X}_n(t))^\top$ is a $(n, p\times L)$ matrix such that its i-th row is given by

$$\boldsymbol{X}_i(t) = \left((X_i^1(t)\varphi_1(t),\dots,X_i^1(t)\varphi_L(t)),\dots,(X_i^p(t)\varphi_1(t),\dots,X_i^p(t)\varphi_L(t))\right)^\top$$

Then, one has

$$\ell_n^{\mathrm{A}}(\beta) = \beta^\top \boldsymbol{G}_n^{\mathrm{A}}\beta - 2\beta^\top \nu_n,$$

where $\boldsymbol{G}_n^{\mathrm{A}}$ is $(p\times L, p\times L)$ matrix defined by

$$\boldsymbol{G}_n^{\mathrm{A}} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\boldsymbol{X}_i(t)(\boldsymbol{X}_i(t))^\top dt, \tag{4.18}$$

and $\nu_n \in \mathbb{R}^{p\times L}$ has coordinates

$$(\nu_n)_{j,l} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau X_i^j(t)\varphi_l(t)dN_i(t).$$

Recall that

$$\hat\beta^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}}\left\{\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat\gamma}\right\}, \tag{4.19}$$

Equation (4.19) is equivalent to the following:

$$\hat\beta^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}}\left\{\ell_n^{\mathrm{A}}(\beta) + \|\hat\gamma\odot D\beta\|_1\right\},$$

D is a p block diagonal matrix, i.e., $D = \mathrm{diag}(D_1,\dots,D_p)$, such that for $j = 1,\dots,p, D_j$ is a (L,L) matrix given by

$$D_j = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0\\ -1 & 1 & 0 & \cdots & 0\\ 0 & -1 & 1 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & -1 & 1\end{bmatrix}.$$

Now let us consider the $(p\times L)$-dimensional vector $\mu = (\mu_{1,1},\dots,\mu_{1,L},\dots,\mu_{p,1},\dots,\mu_{p,L})^\top$ defined by the relation $\beta = T\mu$, where T is an operator given by by the p block diagonal matrix, i.e., $T = \mathrm{diag}(T_1,\dots,T_p)$ and for $j = 1,\dots,p, T_j$ is a $(L\times L)$ matrix given by

$$T_j = \begin{bmatrix} 1 & 0 & \cdots & 0\\ 1 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \cdots & 1\end{bmatrix}.$$

Hence, using the fact that $D^{-1} = T$, we can rewrite the estimator $\hat\beta = T\hat\mu$, where $\hat\mu$ is defined by

$$\hat\mu^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}}\left\{\boldsymbol{Q}_n^{\mathrm{A}}(\mu) + \|\hat\gamma\odot\mu\|_1\right\},$$

where

$$\boldsymbol{Q}_n^{\mathrm{A}}(\mu) = \mu^\top T^\top \boldsymbol{G}_n^{\mathrm{A}} T\mu - 2\mu^\top T^\top \nu_n,$$

Using the Doob-Meyer decomposition we get

$$\nu_n = \xi_n + Z_n \tag{4.20}$$

where

$$\xi_n = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \boldsymbol{X}_i(t)\lambda_\star^{\mathrm{A}}(t, X_i(t))Y_i(t)dt \text{ and } Z_n = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \boldsymbol{X}_i(t)dM_i(t).$$

We have,

$$\begin{aligned}
\beta^\top \xi_n &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \beta^\top U^\top(t) X_i(t)\lambda_\star^{\mathrm{A}}(t, X_i(t))Y_i(t)dt \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^\tau X_i^\top(t) U(t)\beta\lambda_\star^{\mathrm{A}}(t, X_i(t))Y_i(t)dt \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \lambda_\beta^{\mathrm{A}}(t, X_i(t))\lambda_\star^{\mathrm{A}}(t, X_i(t))Y_i(t)dt.
\end{aligned}$$

Then, we have

$$\beta^\top \xi_n = \langle \lambda_\beta^{\mathrm{A}}, \lambda_\star^{\mathrm{A}} \rangle_n. \tag{4.21}$$

Therefore,

$$\ell_n^{\mathrm{A}}(\beta) = \|\lambda_\beta^{\mathrm{A}}\|_n^2 - 2\langle \lambda_\beta^{\mathrm{A}}, \lambda_\star^{\mathrm{A}} \rangle_n - 2\beta^\top Z_n,$$

Consequently, for any $\beta \in \mathbb{R}^{p\times L}$, the following holds

$$\ell_n^{\mathrm{A}}(\hat{\beta}^{\mathrm{A}}) - \ell_n^{\mathrm{A}}(\beta) = \|\lambda_{\hat{\beta}^{\mathrm{A}}}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 - \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2(\beta - \hat{\beta}^{\mathrm{A}})^\top Z_n$$

By the definition of $\hat{\beta}^{\mathrm{A}}$, we have

$$\ell_n^{\mathrm{A}}(\hat{\beta}^{\mathrm{A}}) + \|\hat{\beta}^{\mathrm{A}}\|_{\mathrm{gTV},\hat{\gamma}} \le \ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}},$$

for any $\beta \in \mathbb{R}^{p\times L}$, then

$$\begin{aligned}
\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 &\le \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2(\hat{\beta}^{\mathrm{A}} - \beta)^\top Z_n + (\|\beta\|_{\mathrm{gTV},\hat{\gamma}} - \|\hat{\beta}^{\mathrm{A}}\|_{\mathrm{gTV},\hat{\gamma}}) \\
&\le \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2(T(\hat{\mu}^{\mathrm{A}} - \mu))^\top Z_n + (\|\hat{\gamma}\odot\mu\|_1 - \|\hat{\gamma}\odot\hat{\mu}^{\mathrm{A}}\|_1) \\
&\le \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2\sum_{j=1}^p\sum_{l=1}^L [\hat{\mu}^{\mathrm{A}} - \mu]_{j,l}[T^\top Z_n]_{j,l} + (\|\hat{\gamma}\odot\mu\|_1 - \|\hat{\gamma}\odot\hat{\mu}^{\mathrm{A}}\|_1)
\end{aligned}$$

Let us introduce the event

$$\mathscr{E}_n = \bigcap_{j=1}^p \bigcap_{\ell=1}^L \left\{2\left|[T^\top Z_n]_{j,l}\right| \le \hat{\gamma}_{j,l}\right\}.$$

Therefore, on $\mathscr{E}_n$, we have

$$\begin{aligned}
\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 &\le \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2\|\hat{\gamma}\odot\mu\|_1 \\
&\le \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}}.
\end{aligned}$$

Furthermore, the j-th bloc of $T^\top Z_n$ is given by

$$[T^\top Z_n]_j = \frac{1}{n}\sum_{i=1}^n \int_0^\tau T_j^\top (\boldsymbol{X}(t))_i^j dM_i(t),$$

with l-th coefficient

$$[T^\top Z_n]_{j,l} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau X_i^j(t) \sum_{u=l}^L \varphi_u(t) dM_i(t).$$

Let us calculate the probability of the complementary event of $\mathscr{E}_n$. Using a union bound, we have

$$\begin{aligned}
\mathbb{P}[\mathscr{E}_n^c] &= \mathbb{P}\Big[\bigcup_{j=1}^p \bigcup_{l=1}^L \big|[T^\top Z_n]_{j,l}\big| > \hat{\gamma}_{j,l}/2\Big] \\
&\le \sum_{j=1}^p \sum_{l=1}^L \mathbb{P}\Big[\Big|\frac{1}{n}\sum_{i=1}^n \int_0^\tau X_i^j(t) \sum_{u=l}^L \varphi_u(t) dM_i(t)\Big| > \hat{\gamma}_{j,l}/2\Big] \\
&\le \sum_{j=1}^p \sum_{\ell=1}^L \mathbb{P}\Big[\Big|\frac{1}{n}\sum_{i=1}^n \int_0^\tau X_i^j(t) \sum_{u=l}^L \mathbb{1}(I_u)(t) dM_i(t)\Big| > \sqrt{\tau}\hat{\gamma}_{j,l}/(2\sqrt{L})\Big]
\end{aligned}$$

Using Theorem 3 in Gaïffas and Guilloux (2012) and the choice of $\hat{\gamma}_{\ell,j}$ (see Definition 1 in the supplementary material), we obtain $\mathbb{P}[\mathscr{E}_n^c] \le C_A e^{-x}$, for a some constant $C_A > 0$. □

4.6.2 Proof of Theorem 4.3.2: slow oracle inequality in the time-varying Cox model

Using the Doob-Meyer decomposition, we can easily obtain that

$$\begin{aligned}
\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}) - \ell_n^{\mathrm{M}}(\beta) = K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) - K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) \\
- \frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big(\log \hat{\lambda}^{\mathrm{M}}(t, X_i(t)) - \log \lambda_\beta^{\mathrm{M}}(t, X_i(t))\Big) dM_i(t).
\end{aligned}$$

We can rewrite the last term as $(\hat{\beta}^{\mathrm{M}} - \beta)^T Z_n$ with

$$Z_n = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \boldsymbol{X}_i(t) dM_i(t). \tag{4.22}$$

From the definition of the Lasso estimator (4.12), we have

$$\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}) + ||\hat{\beta}^{\mathrm{M}}||_{\mathrm{gTV},\hat{\gamma}} \le \ell_n^{\mathrm{M}}(\beta) + ||\beta||_{\mathrm{gTV},\hat{\gamma}}, \quad \forall \beta \in \mathbb{R}^{p\times L},$$

which can be rewritten by

$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) \le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + (\hat{\beta}^{\mathrm{M}} - \beta)^T Z_n + \Big(||\beta||_{\mathrm{gTV},\hat{\gamma}} - ||\hat{\beta}^{\mathrm{M}}||_{\mathrm{gTV},\hat{\gamma}}\Big).$$

Using the same notations for the matrix T and the pL-dimensional vector μ than in the proof of Theorem 4.3.1, we have

$$\begin{aligned}
K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) &\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + (\hat{\beta}^{\mathrm{M}} - \beta)^T Z_n + \Big(||\beta||_{\mathrm{gTV},\hat{\gamma}} - ||\hat{\beta}^{\mathrm{M}}||_{\mathrm{gTV},\hat{\gamma}}\Big) \\
&\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + (T(\hat{\mu}^{\mathrm{M}} - \mu))^T Z_n + \Big(||\hat{\gamma} \odot \mu||_1 - ||\hat{\gamma} \odot \hat{\mu}^{\mathrm{M}}||_1\Big) \\
&\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + \sum_{j=1}^p \sum_{l=1}^L (\hat{\mu}^{\mathrm{M}} - \mu))_{j,l} [T^T Z_n]_{j,l} + \Big(||\hat{\gamma} \odot \mu||_1 - ||\hat{\gamma} \odot \hat{\mu}^{\mathrm{M}}||_1\Big).
\end{aligned}$$

Let us introduce the following set

$$\tilde{\mathscr{E}}_n = \bigcap_{j=1}^{p} \bigcap_{l=1}^{L} \{|[T^T Z_n]_{j,l}| \leq \hat{\gamma}_{j,l}\}.$$

On $\tilde{\mathscr{E}}_n$,

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) &\leq K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + ||\hat{\gamma} \odot \mu||_1 \\ &\leq K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2||\beta||_{\mathrm{gTV},\hat{\gamma}}. \end{aligned}$$

It remains to calculate the probability of $\tilde{\mathscr{E}}_n$. We have

$$[T^\top Z_n]_{j,l} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \sum_{k=l}^{L} X_i^j(t) \varphi_k(t) dM_i(t).$$

Using Theorem 3 in Gaïffas and Guilloux (2012), we can find a bound of $\mathbb{P}(\tilde{\mathscr{E}}_n^c)$ associated to a choice of the weights $\hat{\gamma}_{j,l}$ (see Definition 1 in the supplementary material),

$$\begin{aligned} \mathbb{P}[\tilde{\mathscr{E}}_n^c] &= \mathbb{P}\Big[\bigcup_{j=1}^{p} \bigcup_{l=1}^{L} |[T^\top Z_n]_{j,l}| > \hat{\gamma}_{j,l} \Big] \\ &\leq \sum_{j=1}^{p} \sum_{l=1}^{L} \mathbb{P}\Big[\Big| \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau X_i^j(t) \sum_{k=l}^{L} \mathbb{1}_{I_k}(t) dM_i(t) \Big| > \sqrt{\frac{\tau}{L}} \hat{\gamma}_{j,l} \Big] \\ &\leq C_M \mathrm{e}^{-x}. \end{aligned}$$

□

Supplementary Materials for: Time-Varying High-Dimensional Aalen and Cox Models

This is a supplementary materials of Alaya et al. (2016a). It contains further theoretical results, for instance fast oracle inequalities for both time-varying Aalen and Cox models, and change-point detection problem. That is, we work under the assumption that the regression coefficients $\beta_j^\star$ in the time-varying Aalen model can be well approximated by a piecewise constant functions, and we deal with this problem with a signal segmentation point-of-view.

Contents

Notation. We give some notation that will be used frequently in the sequel. Given two sets $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$, we denote by $[A, B]$ the concatenation of A and B, namely $[A, B] = \{a_1, \ldots, a_r, b_1, \ldots, b_s\}$. For any set $S \subset A$, the complement S^c is defined with respect to A, $\mathbb{1}(S)$ stands for the indicator function of S and its cardinality as $|S|$. Given a vector $u \in \mathbb{R}^m$, $[u]_S$ is the projection, in $\mathbb{R}^r$, of u onto S.

The weights $\hat{\gamma}_{j,}$ used in the definition of the $\ell_1 + \ell_1$-total-varaition have an explicit form given by

Definition 4.6.1. (Weights) For $x > 0$, we introduce the data-driven weights

$$\hat{\gamma}_{j,l} = \frac{5.64}{\sqrt{\tau}} \sqrt{\frac{L(x+\log(pL)+\hat{A}_{n,x,j,l})}{n} \hat{V}_{j,l}} + \frac{18.62}{\sqrt{\tau}} \frac{\sqrt{L}(x+\log(pL)+1+\hat{A}_{n,x,j,l})}{n} \|X^j\|_{n,l,\infty}, \tag{4.23}$$

where

$$\hat{A}_{n,x,j,l} = 2\log\log\Big(\frac{2en\hat{V}_{j,l} + 18.66e(x+\log(pL))\|X^j\|_{n,l,\infty}}{8\|X^j\|^2_{n,l,\infty}} \vee e \Big).$$

and

$$\|X^j\|_{n,l,\infty} = \sup_{t \in \bigcup_{u=l}^{L} I_u} \max_{i=1,\dots,n} \|X^j(t)\|_{n,\ell,\infty}.$$

Note that the weights $\hat{\gamma}_{j,l}$ are fully data-driven. The shape of these weights comes comes from a Bernstein's concentration with data-driven variance, necessary for the control of the noise term (a martingale with jumps), see Theorem 3 in Gaïffas and Guilloux (2012).

Remark 4.6.2. We have

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \le \|\beta\|_{\mathrm{gTV},1} \max_{j=1,\dots,p} \max_{l=1,\dots,L} \Big\{ \frac{5.64}{\sqrt{\tau}} \sqrt{\frac{L(x+\log(pL)+\hat{A}_{n,x,\ell,j})}{n} \hat{V}_{j,\ell}} + \frac{18.62}{\sqrt{\tau}} \frac{\sqrt{L}(x+\log(pL)+1+\hat{A}_{n,x,\ell,j})}{n} \|X^j\|_{n,\ell,\infty} \Big\}$$

for any $\beta \in \mathbb{R}^{p \times L}$.

4.7 Fast oracle inequalities

For any $\beta = (\beta_{1,\cdot}^\top, \dots, \beta_{p,\cdot}^\top)^\top \in \mathbb{R}^{p \times L}$, let $\mathscr{A}(\beta) = [\mathscr{A}_1(\beta_{1,\cdot}), \dots, \mathscr{A}_p(\beta_{p,\cdot})]$ be the support of β relative to the weighted $\ell_1 + \ell_1$-total-variation penalization and it is defined as the concatenation of index sets where for all $j = 1, \dots, p$, $\mathscr{A}_j$ is the support of the discrete gradient of β_j, namely,

$$\mathscr{A}_j(\beta_{j,\cdot}) = \{l : \beta_{j,l} \neq \beta_{j,l-1}, \text{ for } l = 2, \dots, L\} \cup \{1 : \beta_{j,1} \neq 0\}, \tag{4.24}$$

and its complementary $\mathscr{A}^c(\beta) = [\mathscr{A}_1^c(\beta_{1,\cdot}), \dots, \mathscr{A}_p^c(\beta_{p,\cdot})]$.

We recall that the Gram matrices $(\boldsymbol{G}_n^{\mathrm{A}})$ (see (4.18)) and $(\boldsymbol{G}_n^{\mathrm{M}})$ (see (4.29)) are symmetrical semidefinite matrices, their squared are $(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}$ and $(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}$ are well-defined. Usually, in order to obtain a fast oracle inequality, we need to assume a Restricted Eigenvalue condition on $(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}$ and $(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}$. However, since they are random in our case, we impose the Restricted Eigenvalue condition to $\mathbb{E}[(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}]$ and $\mathbb{E}[(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}]$, where the expectation is taken conditionally to the covariates. We recall that the matrices

4.7.1 The time-varying Aalen model

Assumption 4.7.1. *Assume the following condition holds*

$$\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta)) = \inf_{u \in \mathbb{R}^{p\times L}\setminus\{0\}: u \in \mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))} \left\{ \frac{\|\mathbb{E}[(\boldsymbol{G}_n^{\mathrm{A}})]^{1/2} u\|_2}{\sqrt{n}\|[u]_{\mathscr{A}}\|_2} \right\} > 0, \tag{4.25}$$

where

$$\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta)) = \left\{ u \in \mathbb{R}^{p\times L} : \|[u]_{\mathscr{A}^c(\beta)}\|_{\mathrm{gTV},\hat{\gamma}} \leq 3\|[u]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat{\gamma}} \right\}. \tag{4.26}$$

The set $\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))$ consists of all vectors that have support similar to the support of β. Note that the assumption made in Bickel et al. (2009) is slightly stronger but only depends on the cardinality of $\mathscr{A}(\beta)$, by minimizing with respect to all sets with cardinality equal to one of $\mathscr{A}$. Additionally, $\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))$ is a cone that involves the weighted total-variation penalization. The quantity $1/\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))$ is a lower bound for eigenvalues restricted over vectors with a support close to the support of β.

Lemma 4.7.2. *Let* $\|\boldsymbol{X}\|_\infty = \sup_{t\in[0,\tau]} \max_{1\leq i\leq n} \max_{1\leq j\leq p} |X_i^j(t)|$, *and* $c(\hat{\gamma}) = \min_{j,l} \hat{\gamma}_{j,l}^2 / \max_{j,l} \hat{\gamma}_{j,l}^2$. *Given Assumption 4.7.1, one has*

$$\kappa^{\mathrm{A}}(\mathscr{A}(\beta)) = \inf_{u \in \mathbb{R}^{p\times L}\setminus\{0\}: u \in \mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))} \left\{ \frac{||(\boldsymbol{G}_n^{\mathrm{A}})^{1/2} u||_2}{\sqrt{n}||[u]_{\mathscr{A}(\beta)}||_2} \right\} > 0, \tag{4.27}$$

and $\kappa^{\mathrm{A}}(\mathscr{A}(\beta)) = (1/\sqrt{2})\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))$ *with probability larger than* $1-\pi_n^{\mathrm{A}}$, *where*

$$\pi_n^{\mathrm{A}} = \exp\Big(- \frac{n^3/((pL)^4) c(\hat{\gamma}) \big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2}{2\tau\|\boldsymbol{X}\|_\infty^2(\tau\|\boldsymbol{X}\|_\infty^2 + c(\hat{\gamma})\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2/(3(pL)^2))} \Big).$$

The proof of Lemma 4.7.2 is postponed in Section 4.9.1. Thanks to Lemma 4.7.2, the empirical Restricted Eigenvalue condition will be fulfilled on an event of large probability, on which we establish a fast non-asymptotic oracle inequality.

Theorem 4.7.3. *Fix* $x > 0$ *and let Assumption 4.7.1 holds. One has with a probability larger than* $1 - C_A e^{-x} - \pi_n^{\mathrm{A}}$

$$\|\lambda_\star^{\mathrm{A}} - \hat{\lambda}^{\mathrm{A}}\|_n^2 \leq \inf_{\beta\in\mathbb{R}^{p\times L}} \Big(\|\lambda_\star^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 + \frac{608\tau|\mathscr{A}(\beta)|}{L\big(\kappa^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2} \max_{j=1,\ldots,p} \|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2 \Big). \tag{4.28}$$

The proof of Theorem 4.7.3 is presented in Section 4.9.2.

4.7.2 The time-varying Cox model

We introduce the following $(pL \times pL)$-matrix for $t \in [0,\tau]$

$$\boldsymbol{G}_n^{\mathrm{M}} = \frac{1}{n}\int_0^\tau (\boldsymbol{X}(t))^\top \boldsymbol{C}(t)\boldsymbol{X}(t)dt, \quad \text{with} \quad \boldsymbol{C} = \Big(\mathrm{diag}(\lambda_\star^{\mathrm{M}}(t, X_i(t))Y_i(t)\Big)_{1\leq i\leq n}. \tag{4.29}$$

The matrix $\boldsymbol{G}_n^{\mathrm{M}}$ depend on the random process Y_i, so that it is random, even conditionnally to the covariates. Thus, we consider a Restricted Eigenvalue condition applied to the matrix $\mathbb{E}[\boldsymbol{G}_n^{\mathrm{M}}]$.

Assumption 4.7.4. *We assume the following condition*

$$\kappa_0^{\mathrm{M}}(\mathscr{A}(\beta)) = \inf_{u\in\mathbb{R}^{p\times L}\setminus\{0\}:u\in\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))}\left\{\frac{||\mathbb{E}[(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}]u||_2}{\sqrt{n}||[u]_{\mathscr{A}(\beta)}||_2}\right\} > 0, \tag{4.30}$$

where $\mathscr{A}(\beta)$ *is defined by* (4.24) *and* $\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))$ *by* (4.26).

Now, we state a lemma that connect the previous Restricted Eigenvalue Condition (4.30) to an empirical one that we could apply directly to $\boldsymbol{G}_n^{\mathrm{M}}$.

Assumption 4.7.5. *We assume that*

$$A_0 = \sup_{1\le i\le n}\left\{\int_0^\tau \lambda_\star^{\mathrm{M}}(t,X_i(t))dt\right\} < \infty.$$

Lemma 4.7.6. *Under Assumptions 4.7.6 and 4.7.5, we have*

$$\kappa^{\mathrm{M}}(\mathscr{A}(\beta)) = \inf_{u\in\mathbb{R}^{p\times L}\setminus\{0\}:u\in\mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))}\left\{\frac{||(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}u||_2}{\sqrt{n}||[u]_{\mathscr{A}(\beta)}||_2}\right\} > 0, \tag{4.31}$$

and $\kappa^{\mathrm{M}}(\mathscr{A}(\beta)) = (1/\sqrt{2})\kappa_0^{\mathrm{M}}(\mathscr{A}(\beta))$ *with probability larger than* $1-\pi_n^{\mathrm{M}}$, *where*

$$\pi_n^{\mathrm{M}} = \exp\Big(-\frac{n^3/((pL)^4)c(\hat{\gamma})\big(\kappa_0^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2}{2A_0\|\boldsymbol{X}\|_\infty^2(A_0\|\boldsymbol{X}\|_\infty^2 + c(\hat{\gamma})\big(\kappa_0^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2/(3(pL)^2))}\Big).$$

Now, we shall work locally on $B_{pL}(R) = \{\beta\in\mathbb{R}^{p\times L} : ||\beta||_2 \le R\}$.

Theorem 4.7.7. *Let* $\zeta > 0$ *and* $x > 0$. *Under Assumptions 4.7.4 and 4.7.5, we have with probability larger than* $1 - C_M\mathrm{e}^{-x} - \pi_n^{\mathrm{M}}$,

$$K_n(\lambda_\star^{\mathrm{M}},\hat{\lambda}^{\mathrm{M}}) \le (1+\zeta)\inf_{\beta\in B_{pL}(R)}\Big(K_n(\lambda_\star^{\mathrm{M}},\lambda_\beta^{\mathrm{M}}) + C(\zeta,\tau)\frac{|\mathscr{A}(\beta)|}{L\big(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2}\max_{j=1,\dots,p}\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2\Big). \tag{4.32}$$

The proof of Theorem 4.7.7 is presented in Section 4.9.3.

Remark 4.7.8. Note that

$$\frac{|\mathscr{A}(\beta)|}{L}\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2 \le 2|\mathscr{A}(\beta)|\max_{l\in\mathscr{A}_j(\beta_{j,\cdot})}\Big\{31.81\frac{(x+\log(pL)+\hat{A}_{n,x,j,l})}{n}\hat{V}_{j,l} + 345.71\Big(\frac{x+\log(pL)+1+\hat{A}_{n,x,j,l}}{n}\|X^j\|_{n,l,\infty}\Big)^2\Big\},$$

so the dominant term is, up to the loglog term, or order $|\mathscr{A}(\beta)|\log(pL)/n$. This is the fast rate to be found in sparse oracles inequalities Alaya et al. (2015); Dalalyan et al. (2014); Gaïffas and Guilloux (2012); Koltchinskii (2009).

4.8 Piecewise constant regression coefficients in the time-varying Aalen model

In this section, we work under the assumption that the regression coefficients $\beta_j^\star$ in the time-varying Aalen model can be well approximated by a piecewise constant functions, and we deal with this problem with a signal segmentation point-of-view.

Assumption 4.8.1. *For all $j = 1,\dots,p$, we assume*

$$\beta_j^\star(t) = \sum_{k=1}^{K_j^\star} \beta_{j,k}^\star \mathbb{1}(J_{j,k}^\star)(t), \tag{4.33}$$

for all $0 \le t \le \tau$ and where $J_{j,k}^\star = (\tau_{j,k-1}^\star, \tau_{j,k}^\star]$ for $k = 1,\dots,K_j^\star$ and $\tau_{j,0}^\star = 0 < \tau_{j,1}^\star < \dots < \tau_{K_j^\star,j}^\star = \tau$.

Assumption 4.8.1 means that $K_j^\star$ changes affect the value of $\beta_j^\star$ at unknown instants $\tau_{k,j}^\star$. The number of change-points $K_j^\star$ is unknown. Therefore, the goal is to find the unknown times of abrupt changes in the dynamic regression of the intensity. This is referred to multiple change-point problem in statistical literature, see Khodadadi and Asgharian (2008) for a recent review with interesting references. A change-point is a time or position where the structure of the object changes and the goal of change-point detection is to estimate these positions. we want to recover $\lambda_\star^{\mathrm{A}}$ by jointly estimating $K_j^\star, \tau_{j,k}^\star$, and $\beta_{j,k}^\star$ for $k = 1,\dots,K_j^\star$.

For all $j = 1,\dots,p$, we define $\beta_j^{\star,\mathscr{H}}$ the orthogonal projection function of $\beta_j^\star$ on $\mathscr{H}_L$ endowed with the Hilbert norm, and $\Delta_j^\star = \max_{1\le k,k'\le K_j^\star} |\beta_{j,k}^\star - \beta_{j,k'}^\star|$ be the maximum jump size of $\beta_j^\star$. Under Assumption 4.8.1, a control of the approximation term leads to the following.

Proposition 4.8.2. *Let Assumption 4.8.1 holds. Under the same assumptions as the ones from Theorem 4.3.1, one has*

$$\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 \le \frac{2\tau}{L} \sup_{t\in[0,\tau]} \max_{i=1,\dots,n} \|X_i(t)\|_2^2 \sum_{j=1}^{p} (K_j^\star - 1)\Delta_j^\star + 2\|\beta^{\star,\mathscr{H}}\|_{\mathrm{gTV},1} \max_{j=1,\dots,p} \max_{l=1,\dots,L} \hat{\gamma}_{j,l}.$$

The proof of Proposition 4.8.2 is postponed in Section 4.9.4. The approximation term can be small but the price to pay may be important which leads to a large values for the variance one. A consequence of Proposition 4.8.2 is that an optimal tradeoff between approximation and complexity is given by the choice $L \approx n^{1/3}$, and $L \approx n^{1/2}$.

Next, following Alaya et al. (2015), we show that the proposed total-variation with data-driven weights procedure is consistent for the estimation of the change-point positions of each j-th bloc, namely $\mathscr{T}_j^\star = \{\tau_{j,1}^\star,\dots,\tau_{j,K_j^\star}^\star\}$, for a fixed $j = 1,\dots,p$.

Assumption 4.8.3. *Under Assumption 4.8.1, assume that there is a positive constant $c_j \ge 8$ such that*

$$\min_{1\le k\le K_j^\star} |\tau_{j,k}^\star - \tau_{j,k-1}^\star| \ge \frac{c_j}{L}.$$

Assumption 4.8.3 guaranties that the weighted $\ell_1 + \ell_1$-total-variation procedure will be able to recover the (unique) intervals I_{j,l_k}, for $k = 0,\dots,K_j^\star$, where the change-point belongs. To proceed, we define the approximate change-points sequence $l_{j,k}$ to be the right-hand side boundary of the unique interval I_{j,l_k}, that contains the change-point $\tau_{j,k}^\star$, i.e.,

$$\tau_{j,k}^\star \in \left(\frac{l_{j,k-1}}{L}\tau, \frac{l_{j,k}}{L}\tau\right]$$

for $k = 1,\dots,K_j^\star$, where we put by convention $l_{j,0} = 0$ and $l_{K_j^\star,j} = L$.

Given the j-th support $\hat{\mathcal{A}}_| = \{\hat{l}_{j,1},\dots,\hat{l}_{j,\widehat{K_j^\star}}\}$ with $\hat{l}_{j,1} < \dots < \hat{l}_{j,\widehat{K_j^\star}}$ of the discrete gradient of $\hat{\beta}_j^{\mathrm{A}}$and introducing $\hat{l}_{j,0} = 0$ and $\hat{l}_{\widehat{K_j^\star}+1,j} = L$, we define simply

$$\hat{\tau}_{j,k} = \frac{\hat{l}_{j,k}}{L}\tau, \tag{4.34}$$

for $k = 1,\dots,\widehat{K_j^\star}$. In order to be able to prove a consistency results for change-points detection, we need a set of assumptions that quantifies the asymptotic interplay between several quantities:

- $\delta_j^\star = \min_{1\le k\le K_j^\star} |l_{k,j} - l_{k-1,j}|$, which is the minimum distance between two consecutive terms in the change-points of $\beta_j^\star$.
- $\Delta_j^{\star,\mathscr{H}} = \min_{1\le l\le L} |\beta_{j,l}^{\star,\mathscr{H}} - \beta_{j,l-1}^{\star,\mathscr{H}}|$, which is the smallest jump size of the projection $\beta_j^{\star,\mathscr{H}}$ of $\beta_j^\star$ onto $\mathscr{H}$.
- $(\varepsilon_n)_{n\ge1}$ a non-increasing and positive sequence that goes to 0 as $n\to\infty$ and such that $L\varepsilon_n \ge 6$ for any $n \ge 1$.

Assumption 4.8.4. *Assume that* $\delta_j^\star, \Delta_j^{\star,\mathscr{H}}$ *and* $(\varepsilon_n)_{n\ge1}$ *satisfy*

$$\frac{\sqrt{nL}\varepsilon_n\Delta^{\star,\mathscr{H}}}{\sqrt{\log L}} \to \infty \text{ and } \frac{\sqrt{n}\delta_j^\star\Delta^{\star,\mathscr{H}}}{\sqrt{L\log L}} \to \infty$$

as $n\to\infty$.

Proposition 4.8.5. *Let Assumptions 4.8.3 and 4.8.4 hold. If* $\widehat{K_j^\star} = K_j^\star - 1$, *the change-points estimators* $\{\hat{\tau}_{j,1},\dots,\hat{\tau}_{j,\widehat{K_j^\star}}\}$ *given by* (4.34) *satisfy:*

$$\mathbb{P}\big[\max_{1\le k\le K_j^\star-1} |\hat{\tau}_{j,k} - \tau_{j,k}^\star| \le \varepsilon_n\big] \to 1 \tag{4.35}$$

as $n\to\infty$.

In Proposition 4.8.5, for each j-th bloc the number of estimated change points is assumed to be equal to the true number of change points. Since this information is not in general available, we need to relax a little bit the statement of the result given in Proposition 4.8.5. Namely, for each j-th bloc we evaluate a non-symmetrized Hausdorff distance $\mathscr{E}(\hat{\mathscr{T}}_j||\mathscr{T}_j^\star)$ of estimated change-points

$$\hat{\mathscr{T}}_j = \{\hat{\tau}_{j,1},\dots,\hat{\tau}_{j,\widehat{K_j^\star}}\}$$

and the set of true change-points

$$\mathscr{T}_j^\star = \{\tau_{j,1}^\star,\dots,\tau_{j,K_j^\star}\}$$

where for two sets A and B, the quantity $\mathscr{E}(A\|B)$ is given by

$$\mathscr{E}(A\|B) = \sup_{b\in B}\inf_{a\in A}|a-b|.$$

Note that we recover the Hausdorff distance between the sets A and B by evaluating $\max(\mathscr{E}(A\|B),\mathscr{E}(B\|A))$.

Proposition 4.8.6. *Under Assumptions 4.8.3 and 4.8.4 and if* $\widehat{K_j^\star} \geq K_j^\star - 1$, *we have*

$$\mathbb{P}\big[\mathscr{E}(\hat{\mathscr{T}}_j \| \mathscr{T}_j^\star) \leq \varepsilon_n\big] \to 1 \tag{4.36}$$

as $n \to \infty$.

Proposition 4.8.6 means that for each j-th bloc the change-point consistency holds for our weighted $\ell_1 + \ell_1$-total-variation procedure whenever the estimated number of change-points is not less than the true one. The proofs of Propositions 4.8.5 and 4.8.6 are given in Alaya et al. (2015). They are build upon some techniques developed in Harchaoui and Lévy-Leduc (2010), based on a careful inspection of the Karush-Kuhn-Tucker (KKT) optimality conditions, see for instance Boyd and Vandenberghe (2004), for the solutions to the convex problems (4.12) and (4.11). They depend also heavily on a data-driven Bernstein's inequality for the control of the martingale errors.

4.9 Proofs

4.9.1 Proof of Lemma 4.7.2

Let consider the event

$$\Omega_{n,\epsilon}^{\mathrm{A}} = \big\{\big|\big(G_n^{\mathrm{A}} - \mathbb{E}[G_n^{\mathrm{A}}]\big)_{m,m'}\big| \leq \epsilon\big\},$$

for all $\{m, m'\} \in \{1, \ldots, pL\}^2$ with a fixed ϵ. Under Assumption 4.7.1, we have that for all $u \in \mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))$

$$\begin{aligned} u^\top \boldsymbol{G}_n^{\mathrm{A}} u &= u^\top (G_n^{\mathrm{A}} - \mathbb{E}[G_n^{\mathrm{A}}])u + u^\top \mathbb{E}[G_n^{\mathrm{A}}]u \\ &\geq u^\top (G_n^{\mathrm{A}} - \mathbb{E}[G_n^{\mathrm{A}}])u + n\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2 \|[u]_{\mathscr{A}(\beta)}\|_2^2. \end{aligned}$$

On the event $\Omega_{n,\epsilon}^{\mathrm{A}}$ and under Assumption 4.7.1, we get

$$\begin{aligned} u^\top \boldsymbol{G}_n^{\mathrm{A}} u &\geq -\sum_{m,m'} \epsilon |u_m||u_{m'}| + n\big(\kappa_0^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2 \|[u]_{\mathscr{A}(\beta)}\|_2^2 \\ &\geq -\epsilon \|u\|_1^2 + n\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2 \|[u]_{\mathscr{A}(\beta)}\|_2^2 \\ &\geq -\frac{\epsilon (pL)^2}{\min_{j,l} \hat{\gamma}_{j,l}^2} \big(2\|[u]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat{\gamma}}^2 + 2\|[u]_{\mathscr{A}^c(\beta)}\|_{\mathrm{gTV},\hat{\gamma}}^2\big) + n\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2 \|[u]_{\mathscr{A}(\beta)}\|_2^2 \\ &\geq -\frac{20\epsilon (pL)^2}{\min_{j,l} \hat{\gamma}_{j,l}^2} \|[u]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat{\gamma}}^2 + n\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2 \|[u]_{\mathscr{A}(\beta)}\|_2^2 \\ &\geq -\Big(\frac{40\epsilon (pL)^2}{n} \times \frac{\max_{j,l} \hat{\gamma}_{j,l}^2}{\min_{j,l} \hat{\gamma}_{j,l}^2} + \big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2\Big) \|[u]_{\mathscr{A}(\beta)}\|_2^2. \end{aligned}$$

It remains to calculate $\mathbb{P}[\Omega_{n,\epsilon}^{\mathrm{A}}]$. The (m, m')-coefficient

$$\big(G_n^{\mathrm{A}} - \mathbb{E}[G_n^{\mathrm{A}}]\big)_{m,m'} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \big(Y_i(t) - \mathbb{E}[Y_i(t)]\big) X_i^m(t) X_i^{m'}(t) dt.$$

Put $\xi_i^{m,m'} = \int_0^\tau \big(Y_i(t) - \mathbb{E}[Y_i(t)]\big) X_i^m(t) X_i^{m'}(t) dt$. Using the facts that $|\xi_i^{m,m'}| \leq \tau \|\boldsymbol{X}\|_\infty^2$ and $\sum_{i=1}^n \mathbb{E}[(\xi_i^{m,m'})^2] \leq n\tau^2 \|\boldsymbol{X}\|_\infty^4$, we apply Bernstein's inequality to get

$$\mathbb{P}\Big[\big|\big(G_n^{\mathrm{A}} - \mathbb{E}[G_n^{\mathrm{A}}]\big)_{m,m'}\big| > \epsilon\Big] \leq 2\exp\Big(-\frac{n^2\epsilon^2}{2(n\tau^2\|\boldsymbol{X}\|_\infty^4 + \tau\|\boldsymbol{X}\|_\infty^2\epsilon/3)}\Big).$$

So the probability of $(\Omega_{n,\epsilon}^{\mathrm{A}})^c$, with $\epsilon = c(\hat{\gamma})\frac{n}{(pL)^2}\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2$ where $c(\hat{\gamma}) = \min_{j,l} \hat{\gamma}_{j,l}^2 / \max_{j,l} \hat{\gamma}_{j,l}^2$ is given by

$$\mathbb{P}[(\Omega_{n,\epsilon}^{\mathrm{A}})^c] \leq 2(pL)^2 \exp\Big(-\frac{n^3/((pL)^4) c(\hat{\gamma})\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2}{2\tau\|\boldsymbol{X}\|_\infty^2(\tau\|\boldsymbol{X}\|_\infty^2 + c(\hat{\gamma})\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2/(3(pL)^2))}\Big)$$

via an union bound and by denoting

$$\begin{aligned}\pi_n^{\mathrm{A}} &= \exp\Big(-\frac{n^3/((pL)^4) c(\hat{\gamma})\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2}{2\tau\|\boldsymbol{X}\|_\infty^2(\tau\|\boldsymbol{X}\|_\infty^2 + c(\hat{\gamma})\big(\kappa_0^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2/(3(pL)^2))}\Big) \\ &\asymp O\big(\exp(-n^3/(pL)^4)\big).\end{aligned}$$

□

4.9.2 Proof of Theorem 4.7.3: fast oracle inequality in the Aalen time-varying model

In this proof, we adopt techniques of convex optimization used in Alaya et al. (2016b). Recall that

$$\hat{\beta}^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}} \big\{\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}\big\}, \tag{4.37}$$

We denote by $\partial\phi$ the subdifferential mapping of a convex function ϕ. The function $\beta \mapsto \ell_n^{\mathrm{A}}(\beta)$ is differentiable, so the subdifferential of $\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}$ at $\beta \in \mathbb{R}^{p\times L}$ is given by

$$\partial(\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}) = \nabla\ell_n^{\mathrm{A}}(\beta) + \partial(\|\beta\|_{\mathrm{gTV},\hat{\gamma}}).$$

In Equation (4.37), $\hat{\beta}^{\mathrm{A}}$ is an optimum of the objective function if an only if there exists a sequence of subgradients $\hat{u}^{\mathrm{A}} = [\hat{u}_{j,l}]_{1\leq l\leq L, 1\leq j\leq p} \in \partial(\|\hat{\beta}^{\mathrm{A}}\|_{\mathrm{gTV},\hat{\gamma}})$ such that

$$\nabla\ell_n^{\mathrm{A}}(\hat{\beta}^{\mathrm{A}}) + \hat{u}^{\mathrm{A}} = 0 \tag{4.38}$$

Recall that for every $\beta \in \mathbb{R}^{p\times L}$, $\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \|\hat{\gamma} \odot D\beta\|_1$, where the matrix D is given in the proof of Theorem 4.3.1. Put $\hat{u}^{\mathrm{A}} = (\hat{u}_{1,\cdot}^\top, \ldots, \hat{u}_{p,\cdot}^\top)^\top$ such that $\hat{u}_{j,\cdot} = (\hat{u}_{j,1}, \ldots, \hat{u}_{j,L})^\top \in \mathbb{R}^L$. Let us define the concatenation of index sets $\hat{\mathscr{A}}(\hat{\beta}^{\mathrm{A}}) = [\hat{\mathscr{A}}_1(\hat{\beta}_{1,\cdot}^{\mathrm{A}}), \ldots, \hat{\mathscr{A}}_p(\hat{\beta}_{p,\cdot}^{\mathrm{A}})]$ where $\hat{\mathscr{A}}_j(\hat{\beta}_{j,\cdot}^{\mathrm{A}})$ is defined as in (4.24) for all $j = 1, \ldots, p$. Let $\hat{\mathscr{A}}^c(\hat{\beta}^{\mathrm{A}}) = [\hat{\mathscr{A}}_1^c(\hat{\beta}_{1,\cdot}^{\mathrm{A}}), \ldots, \hat{\mathscr{A}}_p^c(\hat{\beta}_{p,\cdot}^{\mathrm{A}})]$ the complementary of $\hat{\mathscr{A}}(\hat{\beta}^{\mathrm{A}})$. By subdifferential calculus we have

$$\begin{cases} \hat{u}_{j,l} = \Big[D_j^\top\big(\hat{\gamma}_{j,\cdot} \odot \mathrm{sign}(D_j\hat{\beta}_j)\big)\Big]_l, \text{ if } l \in \hat{\mathscr{A}}_j(\hat{\beta}_{j,\cdot}^{\mathrm{A}}) \\ \hat{u}_{j,l} \in \Big[D_j^\top\big(\hat{\gamma}_{j,\cdot} \odot [-1,+1]^L\big)\Big]_l, \text{ if } l \in \hat{\mathscr{A}}_j^c(\hat{\beta}_{j,\cdot}^{\mathrm{A}}), \end{cases}$$

where for every $v \in \mathbb{R}^m$, $\mathrm{sign}(v) = (\mathrm{sign}(v_1), \ldots, \mathrm{sign}(v_m))^\top$ with $\mathrm{sign}(v_j)$ defined as the subdifferential of the function $v_j \mapsto |v_j|$, that is

$$\mathrm{sign}(v_j) = \begin{cases} \{1\}, & \text{if } v_j > 0, \\ [-1,1], & \text{if } v_j = 0, \\ \{-1\}, & \text{if } v_j < 0. \end{cases}$$

Therefore, Equation (4.38) is equivalent to

$$\langle 2\boldsymbol{G}_n^{\mathrm{A}}\hat{\beta}^{\mathrm{A}} - 2\nu_n + \hat{u}^{\mathrm{A}}, \hat{\beta}^{\mathrm{A}} - \beta\rangle = 0, \text{ for all } \beta \in \mathbb{R}^{p\times L},$$

where ν_n is given in (4.20). Now using the fact that the subdifferentail mapping is monotone (this is an immediate consequence of its definition, see Rockafellar (1970), to say that $\langle \hat{u}^{\mathrm{A}} - u, \hat{\beta}^{\mathrm{A}} - \beta\rangle \geq 0$, for any $u \in \partial(\|\beta\|_{\mathrm{gTV},\hat{\gamma}})$. Then, we get

$$\langle 2\boldsymbol{G}_n^{\mathrm{A}}\hat{\beta}^{\mathrm{A}} - 2\nu_n, \hat{\beta}^{\mathrm{A}} - \beta\rangle \leq -\langle u, \hat{\beta}^{\mathrm{A}} - \beta\rangle.$$

In one hand, the definition of $\boldsymbol{G}_n^{\mathrm{A}}$, see (4.18), yields that $\beta^\top \boldsymbol{G}_n^{\mathrm{A}}\beta' = \langle \lambda_\beta^{\mathrm{A}}, \lambda_{\beta'}^{A}\rangle_n$. In another hand, recall that $\nu_n = \xi_n + Z_n$ where $\beta^\top \xi_n = \langle \lambda_\beta^{\mathrm{A}}, \lambda_\star^{\mathrm{A}}\rangle_n$, see (4.21) in the proof of Theorem 4.3.1. Hence, it implies that

$$\begin{aligned}
\langle 2\boldsymbol{G}_n^{\mathrm{A}}\hat{\beta}^{\mathrm{A}} - 2\nu_n, \hat{\beta}^{\mathrm{A}} - \beta\rangle &= 2(\hat{\beta}^{\mathrm{A}})^\top \boldsymbol{G}_n^{\mathrm{A}}\hat{\beta}^{\mathrm{A}} - 2(\hat{\beta}^{\mathrm{A}})^\top \boldsymbol{G}_n^{\mathrm{A}}\beta - 2\nu_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\langle \hat{\lambda}^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}}\rangle_n - 2\langle \hat{\lambda}^{\mathrm{A}}, \lambda_\beta^{\mathrm{A}}\rangle_n - 2\nu_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\langle \hat{\lambda}^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}}\rangle_n - 2\langle \hat{\lambda}^{\mathrm{A}}, \lambda_\beta^{\mathrm{A}}\rangle_n - 2(\xi_n + Z_n)^\top(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\langle \hat{\lambda}^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}}\rangle_n - 2\langle \hat{\lambda}^{\mathrm{A}}, \lambda_\beta^{\mathrm{A}}\rangle_n - 2\langle \hat{\lambda}^{\mathrm{A}}, \lambda_\star^{\mathrm{A}}\rangle_n + 2\langle \lambda_\beta^{\mathrm{A}}, \lambda_\star^{\mathrm{A}}\rangle_n - 2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}}\rangle_n - 2\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \lambda_\star^{\mathrm{A}}\rangle_n - 2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\rangle_n - 2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta).
\end{aligned}$$

Then

$$2\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\rangle_n \leq 2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) - u^\top(\hat{\beta}^{\mathrm{A}} - \beta)$$

Using the identity $2\langle u, u'\rangle_n = \|u\|_n^2 + \|u'\|_n^2 - \|u - u'\|_n^2$ we get

$$\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \|\hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 \leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) - u^\top(\hat{\beta}^{\mathrm{A}} - \beta).$$

If $\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\rangle_n < 0$, we have $\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 < \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2$ and then $\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 < \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2$ which entails the Theorem, so we assume that $\langle \hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}, \hat{\lambda}^{\mathrm{A}} - \lambda^\star\rangle_n \geq 0$. Notice that the matrix $TD = I_{(p\times L)}$, the identity matrix, one has the following

$$\begin{aligned}
2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) &= 2(T^\top Z_n)^\top D(\hat{\beta}^{\mathrm{A}} - \beta) \\
&= 2\sum_{j=1}^{p}\sum_{l=1}^{L}[T^\top Z_n]_{j,l}[D(\hat{\beta}_{j,\cdot}^{\mathrm{A}} - \beta_{j,\cdot})]_{j,l}
\end{aligned}$$

Thus, on $\mathscr{E}_n$ we have

$$2Z_n{}^\top(\hat{\beta}^{\mathrm{A}} - \beta) \leq \sum_{l=1}^{L}\hat{\gamma}_{j,l}|[D(\hat{\beta}_{j,\cdot}^{\mathrm{A}} - \beta_{j,\cdot})]_{j,l}| = \|\hat{\beta}^{\mathrm{A}} - \beta\|_{\mathrm{gTV},\hat{\gamma}}$$

Due to the fact that the set $\{1,\ldots,L\} = \mathscr{A}_j(\beta_{j,\cdot}) \cup \mathscr{A}_j^c(\beta_{j,\cdot})$, where $\mathscr{A}_j(\beta_{j,\cdot})$ is the support of the discrete gradient of $\beta_{j,\cdot}$, we have

$$-u^\top(\hat{\beta}^{\mathrm{A}} - \beta) = -\sum_{j=1}^{p}\langle [u_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}, [\hat{\beta}_{j,\cdot}^{\mathrm{A}} - \beta_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\rangle - \sum_{j=1}^{p}\langle [u_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}, [\hat{\beta}^{\mathrm{A}} - \beta]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\rangle$$

Since $u \in \partial(\|\beta\|_{\mathrm{gTV},\hat{\gamma}})$ we can choose

$$\begin{cases}
u_{j,l} = 2\left[D_j^\top\left(\hat{\gamma_{j,\cdot}} \odot \mathrm{sign}(D_j\beta_{j,\cdot})\right)\right]_l, \text{ if } l \in \mathscr{A}_j(\beta_{j,\cdot}) \\
u_{j,l} = 2\left[D_j^\top\left(\hat{\gamma_{j,\cdot}} \odot \mathrm{sign}(D_j\hat{\beta_{j,\cdot}})\right)\right]_l = \left[D_j^\top\left(\hat{\gamma_{j,\cdot}} \odot \mathrm{sign}(D_j(\hat{\beta}_j - \beta_{j,\cdot}))\right)\right]_l, \text{ if } l \in \mathscr{A}_j^c(\beta_{j,\cdot}).
\end{cases}$$

Using a triangle inequality and the fact that $\langle \text{sign}(x), x\rangle = \|x\|_1$, imply that

$$\begin{aligned}
-u^\top(\hat\beta^{\mathrm{A}}-\beta) &= -2\sum_{j=1}^p \langle [D_j^\top(\hat\gamma_{j,\cdot}\odot \text{sign}(D_j\beta_{j,\cdot}))]_{\mathscr{A}_j(\beta_{j,\cdot})}, [\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\rangle \\
&\qquad -2\sum_{j=1}^p \langle [D_j^\top(\hat\gamma_{j,\cdot}\odot \text{sign}(D_j(\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}))]_{\mathscr{A}_j^c(\beta_{j,\cdot})}, [\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\rangle \\
&\le 2\sum_{j=1}^p \|[\hat\gamma_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\odot D_j[\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_1 - 2\sum_{j=1}^p \|[\hat\gamma_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\odot D_j[\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\|_1 \\
&\le 2\sum_{j=1}^p \|[\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_{\mathrm{gTV},\hat\gamma_j} - 2\sum_{j=1}^p \|[\hat\beta^{\mathrm{A}}_{j,\cdot}-\beta_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\|_{\mathrm{gTV},\hat\gamma_j}.
\end{aligned}$$

Therefore, we have

$$\|\hat\lambda^{\mathrm{A}}-\lambda^{\mathrm{A}}_\star\|_n^2 + \|\hat\lambda^{\mathrm{A}}-\lambda^{\mathrm{A}}_\beta\|_n^2 \le \|\lambda^{\mathrm{A}}_\beta-\lambda^{\mathrm{A}}_\star\|_n^2 + 3\|[\hat\beta^{\mathrm{A}}-\beta]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat\gamma}$$

Obviously, we have

$$\|[\hat\beta^{\mathrm{A}}-\beta]_{\mathscr{A}^c(\beta)}\|_{\mathrm{gTV},\hat\gamma} \le 3\|[\hat\beta^{\mathrm{A}}-\beta]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat\gamma},$$

which means that $\hat\beta^{\mathrm{A}}-\beta \in \mathscr{C}_{\mathrm{gTV},\hat\gamma}(\mathscr{A}(\beta))$, see (4.26). Now, in order to control the term $\|[\hat\beta^{\mathrm{A}}-\beta]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat\gamma}$ we use Lemmas 2 and 3 from Alaya et al. (2016b) that give a compatibility conditions satisfied by the matrices T, and $\boldsymbol{G}_n^{\mathrm{A}}T$, respectively, see Lemmas 4.9.1 and 4.9.2 below.

Lemma 4.9.1. *Let $\delta = (\delta_{1,\cdot}^\top,\dots,\delta_{p,\cdot}^\top)^\top \in \mathbb{R}_+^{p\times L}$ be a given vector of "weights". For every $\beta\in\mathbb{R}^{p\times L}$ with a support $\mathscr{A}(\beta)=[\mathscr{A}_1(\beta_{1,\cdot}),\dots,\mathscr{A}_p(\beta_{p,\cdot})]$, such that for all $j=1,\dots,p$, $\mathscr{A}_j(\beta_{j,\cdot})$ is defined like in (4.24), and any $u\in\mathbb{R}^{p\times L}\setminus\{0\}$, we have*

$$\frac{\|Tu\|_2}{\Big|\|[u]_{\mathscr{A}(\beta)}\odot[\delta]_{\mathscr{A}(\beta)}\|_1 - \|[u]_{\mathscr{A}^c(\beta)}\odot[\delta]_{\mathscr{A}^c(\beta)}\|_1\Big|} \ge \varrho_{T,\delta}(\mathscr{A}(\beta)), \tag{4.39}$$

where

$$\varrho_{T,\delta}(\mathscr{A}(\beta)) = \Big\{32\sum_{j=1}^p\sum_{l=1}^L |\delta_{j,l+1}-\delta_{j,l}|^2 + (b_j+1)\|\delta_{j,\cdot}\|_\infty^2\Delta^{-1}_{\min,\mathscr{A}_j(\beta_{j,\cdot})}\Big\}^{-1/2},$$

and $\Delta_{\min,\mathscr{A}_j(\beta_{j,\cdot})} = \min_{1\le r\le b_j}|l_{j,r}-l_{j,r}|$. Here we set $\mathscr{A}_j(\beta_{j,\cdot}) = \{l_{j,1},\dots,l_{j,b_j}\}\subset\{1,\dots,L\}$ with the convention that $l_{j,0}=1$, and $\ell_{j,b_j+1}=L+1$.

Lemma 4.9.2. *Let Assumption 4.7.1 holds. Let $\delta\in\mathbb{R}_+^{p\times L}$, be a given vector of "weights", and $\beta\in\mathbb{R}^{p\times L}$ with a support $\mathscr{A}(\beta)=[\mathscr{A}_1(\beta_{1,\cdot}),\dots,\mathscr{A}_p(\beta_{p,\cdot})]$, such that for all $\mathscr{A}_j(\beta_{j,\cdot})$ is is defined like in (4.24). Then, we have with a probability larger than $1-\pi_n^{\mathrm{A}}$*

$$\begin{aligned}
\inf_{u\in\mathbb{R}^{p\times L}\setminus\{0\}:\,u\in\mathscr{C}_{1,\hat\gamma}(\mathscr{A}(\beta))}\Big\{\frac{\|(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}Tu\|_2}{\sqrt{n}\big|\|[u]_{\mathscr{A}}\odot[\delta]_{\mathscr{A}}(\beta)\|_1 - \|[u]_{\mathscr{A}^c(\beta)}\odot[\delta]_{\mathscr{A}^c(\beta)}\|_1\big|}\Big\} \\
\ge \varrho_{T,\delta}(\mathscr{A}(\beta))\kappa^{\mathrm{A}}(\mathscr{A}(\beta)),
\end{aligned} \tag{4.40}$$

where

$$\mathscr{C}_{1,\hat\gamma}(\mathscr{A}(\beta)) = \Big\{u\in\mathbb{R}^{p\times L}: \sum_{j=1}^p \|[u_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}\|_{1,\gamma_{j,\cdot}} \le 3\sum_{j=1}^p \|[u_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_{1,\gamma_{j,\cdot}}\Big\}, \tag{4.41}$$

with $\|\cdot\|_{1,a}$ denotes the weighted ℓ_1-norm, i.e. $\|v\|_{1,a}=\sum_{j=1}^m a_j|v_j|$, for $(v,a)\in(\mathbb{R}^m\times\mathbb{R}_+^m)$.

Define $\hat{\delta} = (\hat{\delta}_{j,l})_{1\leq l\leq L, 1\leq j\leq p}$ such that

$$\forall j \in [p], \hat{\delta}_{j,l} = \begin{cases} 3\hat{\gamma}_{j,l}, & \text{if } l \in \mathscr{A}_j(\beta_{j,\cdot}), \\ 0, & \text{if } l \in \mathscr{A}_j^c(\beta_{j,\cdot}), \end{cases}$$

Consequently, using Lemmas 4.9.1 and 4.9.2 we get

$$\begin{aligned}
\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \|\hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 &\leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 3\|[\hat{\beta}^{\mathrm{A}} - \beta]_{\mathscr{A}(\beta)}\|_{\mathrm{gTV},\hat{\gamma}} \\
&\leq \|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + 2\frac{\|(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}(\hat{\beta}^{\mathrm{A}} - \beta)\|_2}{\sqrt{n}\varrho_{T,\hat{\delta}}(\mathscr{A}(\beta))\kappa^{\mathrm{A}}(\mathscr{A}(\beta))} \\
&\leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))\big(\kappa^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2} + \frac{1}{n}\|(\boldsymbol{G}_n^{\mathrm{A}})^{1/2}(\hat{\beta}^{\mathrm{A}} - \beta)\|_2^2 \\
&\leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))\big(\kappa^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2} + \|\hat{\lambda}^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2.
\end{aligned}$$

Hence, it follows that

$$\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 \leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A})\big(\kappa^{\mathrm{A}}(\mathscr{A}(\beta))\big)^2}.$$

To finish the proof, it sufficient to find an upper bound for $\frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))}$ which is given by

$$\frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))} = 32\sum_{j=1}^{p}\sum_{l=1}^{L}|\hat{\delta}_{j,l+1} - \hat{\delta}_{j,l}|^2 + 2|\mathscr{A}_j(\beta_{j,\cdot})|\|\hat{\delta}_{j,\cdot}\|_\infty^2\Delta_{\min,\mathscr{A}_j(\beta_{j,\cdot})}^{-1}.$$

Note that $\|\hat{\delta}_{j,\cdot}\|_\infty \leq 3\|\hat{\gamma}_{j,\cdot}\|_\infty$. We write the set $\mathscr{A}_j(\beta_{j,\cdot}) = \{l_{j,1}, ..., l_{j,|\mathscr{A}_j(\beta_{j,\cdot})|}\} \subset \{1,\dots,L\}$ and we set $B_r = [\![l_{j,r-1}, l_{j,r}[\![= \{l_{j,r-1}, l_{j,r-1}+1, \dots, l_{j,r}-1\}$ for $r = 1,\dots,|\mathscr{A}_j(\beta_{j,\cdot})|+1$ with the convention that $l_{j,0} = 1$, and $l_{j,|\mathscr{A}_j(\beta_{j,\cdot})|+1} = L+1$. Then

$$\begin{aligned}
\sum_{l=1}^{L}|\hat{\delta}_{j,l+1} - \hat{\delta}_{j,l}|^2 &= \sum_{r=1}^{|\mathscr{A}_j(\beta_{j,\cdot})|+1}\sum_{l\in B_r}|\hat{\delta}_{j,l+1} - \hat{\delta}_{j,l}|^2 \\
&= \sum_{r=1}^{|\mathscr{A}_j(\beta_{j,\cdot})|+1}\Big\{|\hat{\delta}_{j,l_{j,r-1}+1} - \hat{\delta}_{j,l_{j,r-1}}|^2 + |\hat{\delta}_{j,l_{j,r}} - \hat{\delta}_{j,l_{j,r-1}}|^2\Big\} \\
&= \sum_{r=1}^{|\mathscr{A}_j(\beta_{j,\cdot})|+1}\Big\{\hat{\delta}_{j,l_{j,r-1}}^2 + \hat{\delta}_{j,l_{j,r}}^2\Big\} \\
&= \sum_{r=1}^{|\mathscr{A}_j(\beta_{j,\cdot})|+1} 2\hat{\delta}_{j,l_{j,r}}^2 \\
&\leq 18|\mathscr{A}_j(\beta_{j,\cdot})|\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))} &\leq 32\sum_{j=1}^{p}\Big\{18|\mathscr{A}_j(\beta_{j,\cdot})|\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2\Big\} + 18|\mathscr{A}_j(\beta_{j,\cdot})|\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2\Delta_{\min,\mathscr{A}_j(\beta_{j,\cdot})}^{-1} \\
&\leq 32\sum_{j=1}^{p}\Big\{18 + \frac{1}{\Delta_{\min,\mathscr{A}_j(\beta_{j,\cdot})}}\Big\}|\mathscr{A}_j(\beta_{j,\cdot})|\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2 \\
&\leq 608|\mathscr{A}(\beta)|\max_{j=1,\dots,p}\|[\hat{\gamma}_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}\|_\infty^2.
\end{aligned} \tag{4.42}$$

Finally, we obtain

$$\|\hat{\lambda}^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 \leq \|\lambda_\beta^{\mathrm{A}} - \lambda_\star^{\mathrm{A}}\|_n^2 + \frac{608\tau|\mathscr{A}(\beta|}{L\left(\kappa^{\mathrm{A}}(\mathscr{A}(\beta))\right)^2} \max_{j=1,\ldots,p} \max_{l\in\mathscr{A}_j(\beta_{j,\cdot})} \hat{\gamma}_{j,l}^2.$$

□

4.9.3 Proof of Theorem 4.7.7: fast oracle inequality in the Cox time-varying model

In this proof, we also apply techniques of convex optimization as in the proof of Theorem 4.7.3 and we use the same notations. The function $\beta \mapsto \ell_n^{\mathrm{M}}(\beta)$ is differentiable, so the subdifferential of $\ell_n^{\mathrm{M}}(\beta) + ||\beta||_{\mathrm{gTV},\hat{\gamma}}$ at $\beta \in \mathbb{R}^{pL}$ is given by

$$\partial(\ell_n^{\mathrm{M}}(\beta) + ||\beta||_{\mathrm{gTV},\hat{\gamma}}) = \nabla\ell_n^{\mathrm{M}}(\beta) + \partial(||\beta||_{\mathrm{gTV},\hat{\gamma}}).$$

In Equation (4.12), $\hat{\beta}^{\mathrm{M}}$ is an optimum of the objective function if and only if there exists a sequence of subgradients $\hat{u}^{\mathrm{M}} = [\hat{u}_{j,l}]_{1\leq l\leq L, 1\leq j\leq p} \in \partial(||\hat{\beta}^{\mathrm{M}}||_{\mathrm{gTV},\hat{\gamma}})$ and $\hat{v}^{\mathrm{M}} = [\hat{v}_{j,l}]_{1\leq l\leq L, 1\leq j\leq p} \in \partial(\delta_{B_{pL}(R)}(\hat{\beta}^{\mathrm{M}}))$ such that

$$\nabla\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}) + \hat{u}^{\mathrm{M}} + \hat{v}^{\mathrm{M}} = 0. \tag{4.43}$$

As in proof of Theorem 4.7.3, for $\hat{u}^{\mathrm{M}} = (\hat{u}_{1,1},\ldots,\hat{u}_{1,L},\ldots,\hat{u}_{p,1},\ldots,\hat{u}_{p,L})^\top \in \mathbb{R}^{pL}$, we have

$$\begin{cases} \hat{u}_{j,l} &= \left[D_j^\top(\hat{\gamma}_{j,\cdot} \odot \mathrm{sign}(D_j\hat{\beta}_j))\right]_l, \text{ if } l \in \hat{\mathscr{A}}_j(\beta_{j,\cdot}) \\ \hat{u}_{j,l} &\in \left[D_j^\top(\hat{\gamma}_{j,\cdot} \odot [-1,1]^L)\right]_l, \text{ if } l \in \hat{\mathscr{A}}_j^c(\beta_{j,\cdot}) \end{cases}$$

Equation (4.43) is then equivalent to

$$\langle \nabla\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}) + \hat{u}^{\mathrm{M}} + \hat{v}^{\mathrm{M}}, \hat{\beta}^{\mathrm{M}} - \beta\rangle = 0, \text{ for all } \beta \in \mathbb{R}^{pL}$$

Since the subdifferential mapping is monotone, $\langle \hat{u}^{\mathrm{M}} - u, \hat{\beta}^{\mathrm{M}} - \beta\rangle \geq 0$, for any $u \in \partial(||\beta||_{\mathrm{gTV},\hat{\gamma}})$ and

$$\langle \nabla\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}), \hat{\beta}^{\mathrm{M}} - \beta\rangle \leq -\langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle - \langle \hat{v}^{\mathrm{M}}, \hat{\beta}^{\mathrm{M}} - \beta\rangle.$$

For any $\beta \in B_{pL}(R)$, we have $-\langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle \leq 0$ and then for any $u \in \partial(||\beta||_{\mathrm{gTV},\hat{\gamma}})$

$$\langle \nabla\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}), \hat{\beta}^{\mathrm{M}} - \beta\rangle \leq -\langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle. \tag{4.44}$$

Then, we have

$$\begin{aligned}\langle \nabla\ell_n^{\mathrm{M}}(\hat{\beta}^{\mathrm{M}}), \hat{\beta}^{\mathrm{M}} - \beta\rangle = \Big\langle \frac{1}{n}\sum_{i=1}^n \int_0^\tau \boldsymbol{X}_i(t)Y_i(t)\Big(\lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}}(t,X_i(t)) - \lambda_\star^{\mathrm{M}}(t,X_i(t))\Big)dt, \hat{\beta}^{\mathrm{M}} - \beta\Big\rangle \\ - (\beta^{\mathrm{M}} - \beta)^\top Z_n,\end{aligned}$$

where Z_n is defined by (4.22).

Now, we will find a lower bound for the first scalar product in the right side of the previous equation. For a fixed $\eta \in B_{pL}(R)$, we consider the function $H_n : \mathbb{R} \to \mathbb{R}$, defined by

$$H_n(s) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \exp(\boldsymbol{X_i}(t)^\top(\hat{\beta}^{\mathrm{M}} + s\eta))Y_i(t)dt - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \lambda_\star^{\mathrm{M}}(t,X_i(t))Y_i(t)\langle \boldsymbol{X_i}(t), \hat{\beta}^{\mathrm{M}} + s\eta\rangle dt$$

By differentiating H_n with respect to the variable s we obtain:

$$\dot{H}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{X}_i(t)^\top \eta \exp(\boldsymbol{X}_i(t)^\top(\hat{\beta}^{\mathrm{M}} + s\eta))Y_i(t)dt - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \lambda_\star^{\mathrm{M}}(t, X_i(t))Y_i(t)\boldsymbol{X}_i(t)^\top \eta dt$$
$$\ddot{H}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\boldsymbol{X}_i(t)^\top \eta)^2 \exp(\boldsymbol{X}_i(t)^\top(\hat{\beta}^{\mathrm{M}} + s\eta))Y_i(t)dt$$
$$\dddot{H}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\boldsymbol{X}_i(t)^\top \eta)^3 \exp(\boldsymbol{X}_i(t)^\top(\hat{\beta}^{\mathrm{M}} + s\eta))Y_i(t)dt.$$

We have

$$|\dddot{H}_n(s)| \leq R||\boldsymbol{X}||_\infty |\ddot{H}_n(s)|,$$

where $||\boldsymbol{X}||_\infty = \sup_{t\in[0,\tau]} \max_{1\leq i\leq n} \max_{1\leq j\leq p} |X_i^j(t)|$. We use Lemma 1 in Bach Bach (2010), that we recall here.

Lemma 4.9.3. *Let g be a convex three times differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|g'''(t)| \leq S g''(t)$, for some $S \geq 0$. Then, for all $t \geq 0$:*

$$\frac{g''(0)}{S^2}\psi(St) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}\psi(-St) \text{ with } \psi(u) = \mathrm{e}^{-u} + u - 1$$

Applying Lemma 4.9.3 to H_n, we obtain

$$\ddot{H}_n(0)\frac{\psi(-R||\boldsymbol{X}||_\infty)}{R^2||\boldsymbol{X}||_\infty^2} \leq H_n(s) - H_n(0) - s\dot{H}_n(0),$$

which is equivalent to

$$\ddot{H}_n(0)\frac{\psi(-R||\boldsymbol{X}||_\infty)}{R^2||\boldsymbol{X}||_\infty^2} \leq K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) - K_n(\lambda_\star^{\mathrm{M}}, \lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}}) + \left\langle \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{X}_i(t)Y_i(t)\Big(\lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}}(t, X_i(t)) - \lambda_\star^{\mathrm{M}}(t, X_i(t))\Big)dt, \hat{\beta}^{\mathrm{M}} - \beta \right\rangle$$

From Equation (4.44) and the fact that

$$\ddot{H}_n(0)\frac{\psi(-R||\boldsymbol{X}||_\infty)}{R^2||\boldsymbol{X}||_\infty^2} \geq 0,$$

we deduce

$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}}) \leq K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + \langle Z_n, \hat{\beta}^{\mathrm{M}} - \beta\rangle - \langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle. \tag{4.45}$$

If $\langle Z_n, \hat{\beta}^{\mathrm{M}} - \beta\rangle - \langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle \leq 0$, then Theorem 4.7.7 holds. We assume in the following that

$$\langle Z_n, \hat{\beta}^{\mathrm{M}} - \beta\rangle - \langle u, \hat{\beta}^{\mathrm{M}} - \beta\rangle \geq 0 \tag{4.46}$$

On $\tilde{\mathscr{E}}_n$,

$$\langle Z_n, \hat{\beta}^{\mathrm{M}} - \beta\rangle \leq ||\hat{\beta}^{\mathrm{M}} - \beta||_{\mathrm{gTV},\hat{\gamma}}. \tag{4.47}$$

Since $\{1,\dots,L\} = \mathscr{A}_j(\beta_{j,\cdot}) \cup \mathscr{A}_j^c(\beta_{j,\cdot})$, where $\mathscr{A}_j(\beta_{j,\cdot})$ is the support of the discret gradient of β_j, we have

$$-u^\top(\hat{\beta}^{\mathrm{M}} - \beta) = -\sum_{j=1}^{p} \langle [u_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})}, [\hat{\beta}^{\mathrm{M}}_{j,\cdot} - \beta_{j,\cdot}]_{\mathscr{A}_j(\beta_{j,\cdot})} \rangle - \sum_{j=1}^{p} \langle [u_{j,\cdot}]_{\mathscr{A}_j^c(\beta_{j,\cdot})}, [\hat{\beta}^{\mathrm{M}} - \beta]_{\mathscr{A}_j^c(\beta_{j,\cdot})} \rangle.$$

Since $u \in \partial(||\beta||_{\mathrm{gTV},\hat{\gamma}})$, we can choose

$$\begin{cases} u_{j,l} &= 2\left[D_j^\top(\hat{\gamma}_{j,\cdot} \odot \mathrm{sign}(D_j\hat{\beta}_{j,\cdot}))\right]_l, \text{ if } l \in \hat{\mathscr{A}}_j(\beta_{j,\cdot}) \\ u_{j,l} &= 2\left[D_j^\top(\hat{\gamma}_{j,\cdot} \odot \mathrm{sign}(D_j\hat{\beta}_{j,\cdot}))\right]_l = \left[D_j^\top(\hat{\gamma}_{j,\cdot} \odot \mathrm{sign}(D_j(\hat{\beta}_{j,\cdot} - \beta_{j,\cdot}))\right]_l, \text{ if } l \in \hat{\mathscr{A}}_j^c(\beta_{j,\cdot}). \end{cases}$$

With this choice, we have

$$-u^\top(\hat{\beta}^{\mathrm{M}} - \beta) \le -2||(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}(\beta)}||_{\mathrm{gTV},\hat{\gamma}} - 2||(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}^c(\beta)}||_{\mathrm{gTV},\hat{\gamma}}. \tag{4.48}$$

From Equations (4.46), (4.47) and (4.48), we obtain

$$||(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}^c(\beta)}||_{\mathrm{gTV},\hat{\gamma}} \le 3||(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}(\beta)}||_{\mathrm{gTV},\hat{\gamma}}$$

and

$$\sum_{j=1}^{p} ||(\hat{\gamma}_{j,\cdot})_{\mathscr{A}_j^c(\beta_{j,\cdot})} \odot D_j(\hat{\beta}_{j,\cdot} - \beta_{j,\cdot})_{\mathscr{A}_j^c(\beta_{j,\cdot})}||_1 \le 3\sum_{j=1}^{p} ||(\hat{\gamma}_{j,\cdot})_{\mathscr{A}_j(\beta_{j,\cdot})} \odot D_j(\hat{\beta}_{j,\cdot} - \beta_{j,\cdot})_{\mathscr{A}_j(\beta_{j,\cdot})}||_1,$$

so that $\hat{\beta} - \beta \in \mathscr{C}_{\mathrm{gTV},\hat{\gamma}}(\mathscr{A}(\beta))$. On $\tilde{\mathscr{E}}_n$, from Equation (4.45),

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}}) &\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 3||(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}(\beta)}||_{\mathrm{gTV},\hat{\gamma}} \\ &\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 3||[\hat{\gamma}]_{\mathscr{A}(\beta)} \odot D(\hat{\beta}^{\mathrm{M}} - \beta)_{\mathscr{A}(\beta)}||_{1,\hat{\gamma}} \end{aligned}$$

Let us define $\hat{\delta} = (\hat{\delta}_{j,l})_{1\le l\le L, 1\le j\le p}$ such that

$$\forall j \in [p], \hat{\delta}_{j,l} = \begin{cases} 3\hat{\gamma}_{j,l}, & \text{if } l \in \mathscr{A}_j(\beta_{j,\cdot}), \\ 0, & \text{if } l \in \mathscr{A}_j^c(\beta_{j,\cdot}), \end{cases}$$

Let us also consider the following weighted empirical quadratic norm defined for all function h by

$$||h||_{n,\Lambda} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (h(t, X_i(t)))^2 d\Lambda_i(t)},$$

where Λ_i is defined by

$$\Lambda_i(t) = \int_0^t \lambda_\star^{\mathrm{M}}(s, X_i(t)) Y_i(s) ds,$$

From the definition (4.29) of $\boldsymbol{G}_n^{\mathrm{M}}$, we can easily verify that

$$\frac{1}{n}||(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}(\hat{\beta}^{\mathrm{M}} - \beta)||_2^2 = ||\log\lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}} - \log\lambda_\beta^{\mathrm{M}}||_{n,\Lambda}^2.$$

Let us apply Lemmas 4.9.1 and 4.9.2 to $\boldsymbol{G}_n^{\mathrm{M}}$:

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) &\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2\frac{||(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}(\hat{\beta}^{\mathrm{M}} - \beta)||_2}{\sqrt{n}\varrho_{T,\hat{\delta}}(\mathscr{A}(\beta))\kappa^{\mathrm{M}}(\mathscr{A}(\beta))} \\ &\le K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + \frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))\left(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\right)^2} + \frac{1}{n}||(\boldsymbol{G}_n^{\mathrm{M}})^{1/2}(\hat{\beta}^{\mathrm{M}} - \beta)||_2^2 \\ &\le \tilde{K}_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + \frac{1}{\varrho_{T,\hat{\delta}}^2(\mathscr{A}(\beta))\left(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\right)^2} + ||\log\lambda_{\hat{\beta}^{\mathrm{M}}}^{\mathrm{M}} - \log\lambda_\beta^{\mathrm{M}}||_{n,\Lambda}^2. \end{aligned}$$

Decomposing $||\log\lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}} - \log\lambda^{\mathrm{M}}_{\beta}||^2_{n,\Lambda}$ with $\lambda^{\mathrm{M}}_{\star}$ and applying Proposition 6.2 in Lemler (2013), we obtain

$$K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}) \leq K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\beta}) + \frac{1}{\varrho^2_{T,\hat{\delta}}(\mathscr{A}(\beta))\big(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2} + \frac{1}{\rho'}\tilde{K}_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}) + \frac{1}{\rho'}K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\beta}).$$

Finally, on $\tilde{\mathscr{E}}_n$, we obtain

$$K_n(\lambda^{\mathrm{M}}_{\star},\hat{\lambda}^{\mathrm{M}}) \leq \frac{\rho'+1}{\rho'-1}K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\beta}) + \frac{\rho'}{\rho'-1}\frac{1}{\varrho^2_{T,\hat{\delta}}(\mathscr{A}(\beta))\big(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2}.$$

If we introduce $1+\zeta = (\rho'+1)/(\rho'-1)$ and if we apply the control (4.42) to $\varrho_{T,\hat{\delta}}$, we finally get

$$K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}) \leq (1+\zeta)K_n(\lambda^{\mathrm{M}}_{\star},\lambda^{\mathrm{M}}_{\beta}) + C(\zeta)\frac{608\tau|\mathscr{A}(\beta)|}{L\big(\kappa^{\mathrm{M}}(\mathscr{A}(\beta))\big)^2}\max_{1\leq j\leq p}\max_{1\leq l\leq L}\hat{\gamma}^2_{j,l}.$$

□

4.9.4 Proof of Proposition 4.8.2

We know that $\mathscr{H}_L$ is a subspace of the Hilbert space $\mathbb{L}_2([0,\tau],dt)$ endowed by the norm $\|\beta\| = \big(\int_0^\tau \beta^2(t)dt\big)^{1/2}$. Additionally, $\{\varphi_1,\ldots,\varphi_L\}$ forms an orthonormal basis of $\mathscr{H}_L$ with respect to the previous norm. In the following, we set a connection between the empirical norm $\|\lambda^{\mathrm{A}}_{\beta}\|_n$ and the Hilbert norm $\|\beta\|$.

Lemma 4.9.4. *For every $\lambda^{\mathrm{A}}_{\beta} \in \Lambda^{\mathrm{A}}$, one has*

$$\|\lambda^{\mathrm{A}}_{\beta}\|^2_n \leq \sup_{t\in[0,\tau]}\max_{i=1,\ldots,n}\|X_i(t)\|_2^2\sum_{j=0}^{p}\|\beta_j\|^2.$$

Proof of Lemma 4.9.4. Using Cauchy-Schawrz inequality, we get

$$\begin{aligned}\|\lambda^{\mathrm{A}}_{\beta}\|^2_n &\leq \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \|X_i(t)\|_2^2\|\beta(t)\|_2^2 Y_i(t)dt\\ &\leq \frac{1}{n}\sup_{t\in[0,\tau]}\max_{i=1,\ldots,n}\|X_i(t)\|_2^2\int_0^\tau\|\beta(t)\|_2^2\Big(\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\Big)dt.\end{aligned}$$

Using the fact that for all $t\in[0,\tau]$, $\frac{1}{n}\sum_{i=1}^n Y_i(t)\leq 1$, entails

$$\begin{aligned}\|\lambda^{\mathrm{A}}_{\beta}\|^2_n &\leq \sup_{t\in[0,\tau]}\max_{i=1,\ldots,n}\|X_i(t)\|_2^2\sum_{j=1}^{p}\int_0^\tau\beta_j^2(t)dt\\ &\leq \sup_{t\in[0,\tau]}\max_{i=1,\ldots,n}\|X_i(t)\|_2^2\sum_{j=1}^{p}\|\beta_j\|^2.\end{aligned}$$

Recall that $\lambda^{\mathrm{A}}_{\beta^{\star,\mathscr{H}}}(t,X_i(t)) = X_i^\top(t)\beta^{\star,\mathscr{H}}(t)$ where $\beta^{\star,\mathscr{H}}(t) = \big(\beta_1^{\star,\mathscr{H}}(t),\ldots,\beta_p^{\star,\mathscr{H}}(t)\big)^\top$ and where $\beta_j^{\star,\mathscr{H}}$ is the orthogonal projection of $\beta_j^\star$ onto $\mathscr{H}_L$ with respect to the Hilbert norm. It is clear , by its definition, that $\lambda^{\mathrm{A}}_{\beta^{\star,\mathscr{H}}}(t,X(t))$ belongs to the linear space Λ^{A}, which implies that

$$\inf_{\beta\in\mathbb{R}^{p\times L}}\|\lambda^{\mathrm{A}}_{\beta}-\lambda^{\mathrm{A}}_{\star}\|^2_n \leq \|\lambda^{\mathrm{A}}_{\beta^{\star,\mathscr{H}}}-\lambda^{\mathrm{A}}_{\star}\|^2_n.$$

Besides, using Lemma 4.9.4, we get

$$\|\lambda^{\mathrm{A}}_{\beta^\star,\mathscr{H}} - \lambda^{\mathrm{A}}_\star\|_n^2 \leq \sup_{t\in[0,\tau]} \max_{i=1,\dots,n} \|X_i(t)\|_2^2 \sum_{j=1}^p \|\beta_j^{\star,\mathscr{H}} - \beta_j^\star\|^2.$$

Now, by Lemma 1 in Alaya et al. (2015), we have

$$\|\lambda^{\mathrm{A}}_{\beta^\star,\mathscr{H}} - \lambda^{\mathrm{A}}_\star\|_n^2 \leq \frac{2\tau}{L} \sup_{t\in[0,\tau]} \max_{i=1,\dots,n} \|X_i(t)\|_2^2 \sum_{j=1}^p (K_j^\star - 1)\Delta_j^\star.$$

□

Conclusion

In this thesis an emphasis is placed on estimation procedures based on total-variation penalization. We use this techniques in Chapter 2 to study the problem of learning the inhomogeneous intensity of a counting process, under a sparse segmentation assumption. We prove theoretical results for the prediction error, and consistency in the estimation of change-points.

In Chapter 3, we propose *binarsity* as an appropriate regularization for the binarization technique of continuous features. Through a weighted version of binarsity, we study the estimation problem in generalized linear models.

In high dimensional setting, we study in Chapter 4 dynamic regression models of Aalen and Cox with time-varying covariates and coefficients. We introduce a covariate-specific weighted total variation penalization, using data-driven weights that correctly scale the penalization. We investigate the theoretical properties of the proposed estimators by proving oracle inequalities.

The results in this thesis lead to some future directions as follows:

— We would like to investigate the loss induced by maximum likelihood estimation instead of least-squares for the Aalen models and extend our results to this loss.

— It will be of interest to study a multivariate extension of the proposed one dimensional total variation proximal operator algorithm (Algorithm 3 in Chapter 2).

— Regarding Chapter 3, a future work would be to compare numerically the prediction performance of binarsity with others procedures like CART and Random forests (Breiman et al. (1984)).

Liste des Figures

Bibliographie

O. O. Aalen. Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6(4):701–726, 1978. 20, 32

O. O. Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978)*, volume 2 of *Lecture Notes in Statist.*, pages 1–25. Springer, New York-Berlin, 1980. 108, 110

O. O. Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989. 108

O. O. Aalen. Further results on the non-parametric linear regression model in survival analysis. *Statistics in medicine*, 12(17):1569–1588, 1993. 108

Alan Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015. 24, 81

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974. 13

C. M. Alaiz, A. Barbero, and J. R. Dorronsoro. Group fused lasso. In V. Mladenov, P. Koprinkova-Hristova, G. Palm, A. E. P. Villa, B. Appollini, and N. Kasabov, editors, *Artificial Neural Networks and Machine Learning – ICANN 2013*, volume 8131 of *Lecture Notes in Computer Science*, pages 66–73. Springer Berlin Heidelberg, 2013. 15

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *Information Theory, IEEE Transactions on*, 61(9):5148–5171, 2015. 29, 78, 82, 109, 113, 114, 126, 127, 129, 138

M. Z. Alaya, T. Allart, A. Guilloux, and S. Lemler. Time-varying high-dimensional aalen and cox models. *preprint*, 2016a. 107, 123

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Binarsity: features binarization and cuts selection using convex optimization. *preprint*, 2016b. 77, 109, 114, 130, 132

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. 19, 20, 30, 108, 111

A. Antoniadis, G. Grégoire, and G. Nason. Density and hazard rate estimation for right censored data using wavelet methods. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 61(1):63–84, 1999. 21

F. Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010. 88, 98, 135

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012. 16, 39, 82

F. B. Bach. Consistency of the group lasso and multiple kernel learning. Technical report, Journal of Machine Learning Research, 2007. 15

Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2):239–284, 2009. 21

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. 39, 82

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009a. 16, 18, 19, 117

A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on*, 18(11):2419–2434, 2009b. 16

Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981. 20

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999. 68

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009. 14, 16, 34, 43, 78, 84, 85, 113, 125

L. Birgé. *Model selection for Poisson processes*, volume Volume 55 of *Lecture Notes–Monograph Series*, pages 32–64. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. 21

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. 13

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2007. 13

K. Bleakley and J. P. Vert. The group fused Lasso for multiple change-point detection. 15, 31

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 110

L. Bottou. Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, and K. R. Muller, editors, *Neural Networks: Tricks of the Trade (2nd ed.)*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012. 114

O. Bouaziz and A. Guilloux. A penalized algorithm for event-specific rate models for recurrent events. *Biostatistics*, 16(2):281–294, 2015. 109

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008. 110

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. 18, 23, 37, 40, 92, 129

L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1): 157–183, 2009. 31

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. 79, 139

P. Brémaud. *Point processes and queues*. Springer-Verlag, New York, 1981. Martingale dynamics, Springer Series in Statistics. 19, 32

E. Brunel and F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhya: The Indian Journal of Statistics (2003-2007)*, 67(3): 441–475, 2005. 21

E. Brunel, F. Comte, and C. Lacour. Adaptive estimation of the conditional density in the presence of censoring. *Sankhya: The Indian Journal of Statistics*, pages 734–763, 2007. 21

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications. 14, 78

F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Statist.*, 2:1153–1194, 2008. 15

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, 1:169–194, 2007. 14, 16, 78, 113

Z. Cai and Y. Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003. 108

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007. 14

A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision*, 84(3): 288–307, 2009. 31

A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9:263–340, 2010. 31

O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol.*, 5(4), December 2014. 24, 79

C. Chaux, P. L. Combettes, J. C. Pesquet, and V. R. Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495, 2007. 18

F. Chen, P. F. Yip, and K. F. Lam. On the local polynomial estimators of the counting process intensity function and its derivatives. *Scand. J. Stat.*, 38(4):631–649, 2011. 30

X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.*, 6(2): 719–752, 2012. 16

D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, 2009. 23, 42

B.. Chlebus and S. H. Nguyen. On finding optimal discretizations for two attributes. In Lech Polkowski and Andrzej Skowron, editors, *Rough Sets and Current Trends in Computing*, volume 1424 of *Lecture Notes in Computer Science*, pages 537–544. Springer Berlin Heidelberg, 1998. 79

G. Ciupera. Model selection by lasso methods in a change-point model. *To appear in Statistical Papers*. 31

P. L. Combettes and J. C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011. 16, 18

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200 (electronic), 2005. 16, 18

P.L. Combettes and J. Pesquet. A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):564–574, 2007. 18

F. Comte, S. Gaïffas, and A. Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(4):1171–1196, 2011. 21, 47

L. Condat. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013. 18, 23, 39

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972. 20, 108, 110

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975. 110

D. M. Dabrowska. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, 14(3):181–197, 1987. 20

A. S. Dalalyan, M. Heiri, and J. Lederer. On the prediction performance of the lasso. to appear in Bernoulli 1402.1700, arXiv, February 2014. 93, 109, 126

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 16, 18

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998. 16

J. Dougherty, R. Kohavi, and S. Mehran. Supervised and unsupervised discretization of continuous features. In *Machine Learning: Proceedings of the Twelfth International Conference*, pages 194–202. Morgan Kaufmann, 1995. 23, 79

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12, July 2011. 114

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors. 14

T. Evgeniou, T. Poggio, M. Pontil, and A. Verri. Regularization and statistical learning theory for data analysis. *Comput. Stat. Data Anal.*, 38(4), February 2002. 12

T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991. 19, 117

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007. 15, 18, 78

S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.*, 6:522–546, 2012. 21, 22, 31, 44, 109, 110, 111, 113, 120, 121, 124, 126

S. Garcia, J. Luengo, J. A. Saez, V. Lopez, and F Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4). 79

Melody S. Goodman, Yi Li, and Ram C. Tiwari. Detecting multiple change points in piecewise constant hazard functions. *J. Appl. Stat.*, 38(11):2523–2532, 2011. 31

P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994. 108

G. Grégoire. Least squares cross-validation for counting process intensities. *Scand. J. Statist.*, 20(4):343–360, 1993. 20, 30

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015. 113

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492):1480–1493, 2010. 15, 22, 31, 32, 36, 37, 38, 48, 78, 109, 129

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction. 78

Mohamed Hebiri. *Some variable selection procedures based on the Lasso estimator*. PhD thesis. 14

C. Heuchenne and I. Van Keilegom. Location estimation in nonparametric regression with censored data. *Journal of Multivariate Analysis*, 98(8):1558 – 1582, 2007. 20

D. S. Hochbaum. An efficient algorithm for image segmentation, markov random fields and related problems. *Journal of the ACM (JACM)*, 48(4):686–701, 2001. 31

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 15

T. Honda and W. K. Härdle. Variable selection in cox regression models with varying coefficients. *Journal of Statistical Planning and Inference*, 148:67–81, 2014. 108

J. Huang. Efficient estimation of the partly linear additive cox model. *Ann. Statist.*, 27(5):1536–1563, 10 1999. 20

J. Huang, T. Sun, Z. Ying, Y. Yu, and Zhang C.-H. Oracle inequalities for the lasso in the cox model. *Ann. Statist.*, 41(3):1142–1165, 2013. 21, 109

F. W. Huffer and I. W. McKeague. Weighted least squares estimation for aalen's additive risk model. *Journal of the American Statistical Association*, 86(413):114–129, 1991. 108

L. Jun, Y. Lei, and Y. Jieping. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 323–332, 2010. 15

A.F. Karr. *Point processes and their statistical inference*, volume 7. CRC press, 1991. 20, 32

A. Khodadadi and M. Asgharian. Change-point problems and regression: An annotated bibliography. *Collection of Biostatistics Research Archive (COBRA)*, 2008. 30, 127

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5): 1356–1378, 2000. 14, 78

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15 (3):799–828, 2009. 126

S. Lemler. Oracle inequalities for the lasso in the high-dimensional multiplicative aalen intensity model. *Les Annales de l'Institut Henri Poincaré, arXiv preprint*, 2013. 21, 109, 137

P. A. W Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3). 115

G. Li and H. Doss. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, 23(3):787–823, 06 1995. 20

L. Li, X. Dong, X. Li, and W. Li. Oracle properties of the adaptive elastic net. In *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, volume 3, pages 538–542, Oct 2010. 16

Q. Li and N. Lin. The bayesian elastic net. *Bayesian Anal.*, 5(1):151–170, 03 2010. 15

O. B. Linton, J. P. Nielsen, and S. van de Geer. Estimating multiplicative and additive hazard functions by kernel methods. *The Annals of Statistics*, 31(2):464–492, 2003. 20

R. Sh. Liptser and A. N. Shiryayev. *Theory of martingales*, volume 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. Translated from the Russian by K. Dzjaparidze. 19, 32

M. A. Little and N. S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals. i. background theory. *Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences*, 467:3088–3114, 2011. 15, 78

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: an enabling technique. *Data Min. Knowl. Discov.*, 6(4):393–423, 2002. 79

K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008. 16

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 08 2011. 16

J. Mairal. Stochastic majorization minimization algorithms for large scale optimization. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger,

editors, *Advances in Neural Information Processing Systems 26*, pages 2283–2291. 2013. 113

C. L. Mallows. Some comments on c_p. *Technometrics*, 15(4):661–675, 1973. 13

T. Martinussen and T. H Scheike. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007. 19, 27, 108, 109, 111, 116

T. Martinussen and T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.*, 36(4):602–619, 2009. 21, 109, 110

T. Martinussen, T. H. Scheike, and I. M. Skovgaard. Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics*, 29(1):57–74, 2002. 108

I. W. McKeague. Asymptotic theory for weighted least squares estimators in. In *Statistical Inference from Stochastic Processes: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference Held August 9-15, 1987, with Support from the National Science Foundation and the Army Research Office*, volume 80, page 139. American Mathematical Soc., 1988. 108

I. W. McKeague and K. J. Utikal. Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18(3):1172–1187, 09 1990. 20

Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B*, 70(1):53–71, 2008. 15

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009. 14, 16

P. A. Meyer. A decomposition theorem for supermartingales. *Illinois J. Math.*, (2): 193–205, 06 . 19

J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Paris*, 255:2897–2899, 1962. 17

S. A. Murphy and P. K. Sen. Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180, 1991. 26, 108, 109

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statist.*, 2:605–633, 2008. 15

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995. 13

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972. 25

Y. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London. 17

Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. 16, 18

P. N. Patil and A. T. A. Wood. Counting process intensity estimation by orthogonal wavelet methods. *Bernoulli*, 10(1):1–24, 2004. 20, 30

F. Picard, S. Robin, E. Lebarbier, and J-J Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3):758–766, 2007. 31

J. Qian and L. Su. Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *working paper*, 2013. 31, 32, 38

J. R. Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, January 1993. 79

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11(2):453–466, 1983. 20, 30

F. Rapaport, E. Barillot, and J. P. Vert. Classification of arraycgh data using fused svm. 24(13):i375–i382, 2008. 78

P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1): 103–153, 2003. 21, 22, 30, 31, 32, 34, 44

P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006. 21, 22

P. Reynaud-Bouret and V. Rivoirard. Adaptive thresholding estimation of a poisson intensity with infinite support. *arXiv preprint arXiv:0801.3157*, 2008. 21

A. Rinaldo. Properties and refinements of the fused lasso. *Ann. Statist.*, 37(5B): 2922–2952, 2009. 15, 31, 109

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970. 95, 101, 131

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4), November 1992. 15, 78

T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. *arXiv preprint arXiv:1206.1106*, 2012. 114

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. 13

J. J. Shen and N. R. Zhang. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.*, 6(2):476–496, 2012. 31

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013. 15

W. Stute. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, 23(4):461–471, 1996. 20

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. 13, 78

R. Tibshirani. The lasso method for variable selection in the cox model. *Statist. Med.*, 16:385–395, 1997. 109

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005. 15, 31, 32, 78, 109

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013. 14

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50:2231–2242, 2004. 13

S. Van De Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, 23(5):1779–1801, 1995. 64

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 04 2008. 88

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009. 14, 16, 113

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 12

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. 12

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor.*, 55(5), 2009. 14

A. Winnett and P. Sasieni. Iterated residuals and time-varying covariate effects in cox regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):473–488, 2003. 108, 117

J. Wu and S. Coggeshall. *Foundations of Predictive Analytics (Chapman & Hall / CRC Data Mining and Knowledge Discovery Series)*. 2012. 78

D. Yu, J. Won, T. Lee, J. Lim, and S. Yoon. High-dimensional fused lasso regression using majorization-minimization and parallel processing. *Journal of Computational and Graphical Statistics*, 24(1):121–153, 2015. 31

Y. L. Yu. On decomposing the proximal map. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 91–99. Curran Associates, Inc., 2013. 18, 88

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006. 15

M. D. Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. 114

C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 08 2008. 16

T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 egularization. *Ann. Statist.*, 37(5A):2109–2144, 2009. 14

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7, 2006. 14, 78

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101 (476):1418–1429, 2006. 14, 16

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005. 14, 15

H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, 37(4):1733–1751, 08 2009. 15

D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, pages 329–353, 1990. 108

Résumé

Dans la première partie de cette thèse, nous cherchons à estimer l'intensité d'un processus de comptage par des techniques d'apprentissage statistique en grande dimension. Nous introduisons une procédure d'estimation basée sur la pénalisation par variation totale avec poids. Un premier ensemble de résultats vise à étudier l'intensité sous une hypothèse a priori de segmentation sparse. Dans une seconde partie, nous étudions la technique de binarisation de variables explicatives continues, pour laquelle nous construisons une régularisation spécifique à ce problème. Cette régularisation est intitulée "binarsity", elle compte les valeurs différentes d'un vecteur de paramètres. Dans la troisième partie, nous nous intéressons à la régression dynamique pour les modèles d'Aalen et de Cox avec coefficients et covariables en grande dimension, et pouvant dépendre du temps. Pour chacune des procédures d'estimation proposées, nous démontrons des inégalités oracles non-asymptotiques en prédiction. Nous utilisons enfin des algorithmes proximaux pour résoudre les problèmes convexes sous-jacents, et nous illustrons nos méthodes sur des données simulées et réelles.

Mots-clefs: Processus de comptage, points de rupture, binarisation de variables, régression dynamique, variation-totale, inégalités oracles, algorithmes proximaux

Abstract

In the first part of this thesis, we deal with the problem of learning the inhomogeneous intensity of a counting process, under a sparse segmentation assumption. We introduce a weighted total-variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. In the second part, we study the binarization technique of continuous features, for which we construct a specific regularization. This regularization is called "binarsity", it computes the different values of a parameter. In the third part, we are interested in the dynamic regression models of Aalen and Cox with time-varying covariates and coefficients in high-dimensional settings. For each proposed estimation procedure, we give theoretical guaranties by proving non-asymptotic oracle inequalities in prediction. We finally present proximal algorithms to solve the underlying studied convex problems, and we illustrate our methods with simulated and real datasets.

Keywords: Counting processes, change-points, features binarization, dynamic regression, total-variation, oracle inequalities, proximal algorithms