Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Regressione lineare dinamica penalizzata

Relatore Prof. Mauro Bernardi
Dipartimento di Scienze Statistiche

Correlatore Prof. Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureando Nicolò Zane
Matricola 2039814

Anno Accademico 2022/2023

Indice

Introduzione

L'idea principale per questa tesi nasce dalla curiosità che ho trovato nello studio, nel corso "Modelli statistici per dati ad elevata dimensionalità", del modello noto come *fused-lasso signal-approximator* uno strumento solitamente utilizzato per l'approssimazione di serie storie nell'univariato ed estendibile, in due dimensioni, alla ricostruzione di immagini o matrici. L'obiettivo iniziale quindi è stato quello di elaborare un modello che sfruttasse la struttura di base del *fused-lasso* per riuscire ad affrontare problemi più complessi ed in particolare la regressione dinamica. La tesi inizia con la descrizione di quelli che vengono comunemente chiamati modelli gerarchici per poi generalizzarli e renderli una struttura a più ampio spettro adattabile a molteplici modelli di regressione, dalle *splines* ai modelli *state-space*. Questa connessione permette, dopo aver riscritto il modello in una forma vincolata, di esplicitare il modello *state-space* come un particolare modello ad effetti misti (gerarchico) tramite la costruzione di un'equazione che per mezzo di una matrice, simile a quella utilizzata nel *fused-lasso*, mette a sistema le varie equazioni di stato. Elemento centrale di tutti i modelli che si vedranno durante l'elaborato è rappresentato proprio dalla matrice $\mathbf{F}$ definita anche come matrice delle differenze prime che definisce la distanza tra due parametri adiacenti, in questo caso temporalmente. Prendendo quindi ispirazione dalla forma utilizzata per la regressione *fused-ridge* (Goeman, 2008) viene proposto un modello utilizzabile in contesti dinamici inserendo diverse penalità, inizialmente per mimare la dinamica markoviana dei coefficienti presente nei modelli *state-space* e successivamente per inserire sparsità nel modello come accade nei modelli *LASSO*. Si vedranno in sostanza tre modelli a complessità crescente: il primo denominato *fused-ridge* dinamica, costruita sulla base della regressione ridge generalizzata, il secondo denominato *fused-elastic-net* che aggiunge al modello precedente una penalità norma l_1 per accentuare ancor di più l'effetto della matrice $\mathbf{F}$ e come terzo un modello che aggiunge due penalità *group-LASSO* per introdurre sparsità nel modello. Di questi, si confronteranno i primi due con il filtro di Kalman e il *Kalman-smoother* in uno studio di simulazione, mentre per quanto riguarda l'ultimo,

data la complessità, si esploreranno gli effetti delle due penalità sia singolarmente che congiuntamente. Per stimare questi modelli, che presenteranno sempre una struttura ad elevata dimensionalità ($p > n$), si utilizzerà l'algoritmo chiamato "Metodo alternato dei moltiplicatori"($ADMM$) che verrà descritto in diverse specificazioni e adattato a molteplici casi nel corso dell'elaborato, mostrandone la duttilità e l'efficienza. Nonostante, come si vedrà, la stima dei modelli proposti sia precisa ed efficace, trovandoci nel campo dei modelli penalizzati sarà necessario elaborare dei metodi per la scelta del parametro di penalità ottimale. Il problema principale in questo caso riguarda la struttura del modello che impone una struttura diagonale a blocchi alla matrice di disegno $\mathbf{Z} \in \mathbb{R}^{T \times Tp}$ che oltre a risultare in matrici ad elevata dimensionalità ha una forma rigida che non può essere adattata a tecniche come la validazione incrociata. Si utilizzeranno quindi tecniche di selezione del modello *in-sample* nello specifico i criteri di informazione. Questi metodi però perdono di efficacia al crescere di p ed inoltre in questo contesto sono utilizzabili solo in modelli che hanno un metodo di stima lineare, cioè le cui previsioni $\hat{\mathbf{y}}$ possono essere espresse come prodotto tra una certa matrice $\mathbf{S}$ che non dipende da $\mathbf{y}$ e la variabile risposta $\mathbf{y}$, questo per permette di definire una stima dei gradi di libertà utilizzati che, diversamente dai modelli classici, non sono definiti dal numero di parametri stimati. Nel modello di regressione *fused-ridge* dinamico si utilizzarà il criterio di validazione incrociata generalizzata approssimata ($GACV$) mutuando l'impostazione dalla classica regressione *ridge* ma, successivamente, con l'aggiunta di una penalità non differenziabile, nello specifico norma l_1 questa non sara più utilizzabile. Si proporrà quindi, prendendo in prestito dalla teoria "Majorization-minimization", un modo per approssimare la funzione norma l_1 e stimare una matrice di lisciamento per un modello simil *elastic-net*. Si riuscirà quindi a riprodurre i risultati, rispetto alla capacità di selezionare il modello, raggiunti per il modello di regressione *fused-ridge* dinamico. La tesi quindi oltre alla proposta di nuovi modelli per il lisciamento di relazioni tra serie storiche e il conseguente adattamento dell'algoritmo $ADMM$ per la loro stima, approfondisce il problema della stima della complessità del modello in contesti non standard, in cui la struttura del modello non permette l'utilizzo dei classici strumenti radicati in letteratura. Infine si propone l'adattamento dei modelli proposti ad un problema noto nella letteratura macro-economica cioè la stima e la previsione della curva dei tassi di rendimento in cui si vedrà come il modello proposto sia applicabile anche in un contesto previsionale. In quest'ultimo caso si proporra un approccio innovativo ad un problema classico raggiungendo dei risultati interpretativi non banali.

Capitolo 1

Introduzione ai modelli gerarchici

In questo capitolo si presenteranno i modelli gerarchici, partendo da un approccio specifico che ne giustifica l'utilizzo seguendo i concetti espressi da Snijders and Bosker (2000), per poi generalizzarli in una forma ricondicibile ad un vasto insieme di modelli utili in contesti ad elevata dimensionalità con focus particolare verso i modelli lineari dinamici.

1.1 Introduzione

In statistica spesso il modo migliore per campionare le osservazione è farlo in maniera indipendente. Esistono però metodi più efficienti di campionamento, uno di questi è il campionamento multi-stadio in cui la popolazione di interesse viene suddivisa in sottopopolazioni e le estrazioni avvengono in due passi. Supponiamo ad esempio di suddividere l'insieme di calciatori militanti nel campionato di serie A in sottopopolazioni rappresentate dalla squadra in cui militano, in questo modo l'estrazione avverrebbe in due stadi separati: prima la squadra (cosidetta unità a livello macro) e poi il giocatore (cosidetta unità a livello micro). È chiaro che la probabilità di estrarre un calciatore con un certo stipendio annuo è influenzata dalla squadra in cui gioca e quindi dallo stadio antecedente di campionamento. Questo tipo di campionamento è molto utilizzato nelle scienze sociali in quanto la società è organizzata in "sottopopolazioni" (aziende, quartieri, scuole) e quindi raccogliere i dati seguendo questo procedimento comporta una riduzione dei costi. Questo approccio quindi si focalizza sulla presenza di dipendenza tra le osservazioni provenienti da una determinata sottopopolazione che, sostanzialmente, descrive il fatto che queste si distinguano tra loro per certe caratteristiche; ignorare

questo aspetto, come fanno notare Snijders and Bosker (2000), ponendo in atto un'analisi che tenga conto di un unico livello di campionamento può portare a conclusioni errate. Ad esempio, aggregare le unita di livello micro mediandole al livello macro può essere utile se l'analisi si concentra principalmente su quest'ultima, ma è necessario tenere conto del fatto che la credibilità di questo processo dipenderà dalla numerosità delle unità micro presenti per ogni unità macro. L'aggregazione in questo caso può portare ad alcuni errori:

- utilizzare correlazioni tra le unità a livello macro per fare inferenza sulle unità a livello micro.
- Non tener conto del fatto che variabili aggregate a livello macro non fanno riferimento direttamente alle unità a livello micro.

D'altra parte è possibile anche dissaggregare i dati ignorando così le unità di livello macro. Ad esempio, in uno studio longitudinale si potrebbe ignorare il fatto che alcune misurazioni siano state fatte sullo stesso soggetto trattando così ogni osservazione come se fosse indipendente dalle altre, in questo modo la numerosità campionaria verrebbe sovra-stimata. Infine, se i dati sono raccolti direttamente a livello micro è sicuramente corretto procedere in questo modo purchè si tenga conto del fatto che queste, facendo riferimento alla stessa unità di livello macro, possano essere correlate.

1.2 Modelli lineari ad effetti misti

Un primo modello utilizzato per tenere conto di quanto presentato nella sezione precedente è il seguente:

$$y_{ij} = \mu + u_i + \epsilon_{ij}. \tag{1.1}$$

In questo caso y ha due indici: i (da 1 a q) che indica la sottopopolazione di provenienza e j (da 1 a m) che indica la j-esima osservazione all'interno della sottopopolazione. In questa formulazione u_i è indipendente e identicamente distribuito $N(0, \sigma_s^2)$ e ϵ_{ij} è indipendente e identicamente distribuito $N(0, \sigma_\epsilon^2)$ e indipendente da u_i. In questo modo l'osservazione y_{ij} ha due fonti di varianza: una proveniente dalla sottopopolazione i, rappresentata dal termine u_i, e una derivante dalla singola osservazione (i, j). Questo tipo di modello è chiamato gerarchico o, più in generale, ad effetti misti.

Generalizzando, seguendo l'interpretazione data in Hodges (2014), un modello lineare ad effetti misti può essere definito come un modello di regressione :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N(0, \mathbf{R}(\phi_{\mathbf{R}})),\ \boldsymbol{u} \sim N_q(\mathbf{0}, \mathbf{G}(\phi_{\mathbf{G}})), \tag{1.2}$$

in cui:

- $\boldsymbol{y}$ un vettore n-dimensionale;
- $\mathbf{X}$ la matrice di disegno $n \times p$ che rappresentano gli effetti fissi;
- $\boldsymbol{\beta}$, un vettore p-dimensionale di effetti fissi da stimare;
- $\mathbf{Z}$ una matrice di costanti di dimensioni $n \times q$ che rappresenta gli effetti casuali;
- $\boldsymbol{u}$ un vettore q-dimensionale di effetti casuali da stimare;
- $\boldsymbol{\epsilon}$ un termine di errore;
- $\mathbf{G}$ e $\mathbf{R}$ matrici di varianza e covarianza funzione dei parametri $\phi_{\mathbf{R}}$ e $\phi_{\mathbf{G}}$ da stimare.

Spesso $\mathbf{R}$ viene fissato come $\sigma_e^2\mathbf{I}_n$ ma non è necessario.
La novità di questa famiglia di modelli rispetto ai classici modelli lineari con singolo termine di errore $\boldsymbol{\epsilon}$ è quindi il termine $\mathbf{Z}\boldsymbol{u}$ e di conseguenza la sua struttura di varianza $\mathbf{G}(\phi_{\mathbf{G}})$. Adattando questa formulazione al modello (1.1) avremo $n = qm$, $\mathbf{X} = \mathbf{I}_{qm}$, $\boldsymbol{\beta} = \boldsymbol{\mu}$, $\mathbf{Z} = \mathbf{I}_q \otimes \mathbf{1}_m$ che seleziona l'i-esimo elemento del vettore q-dimensionale $\boldsymbol{u}$, ed infine $\mathbf{G}{=}\sigma_s^2\mathbf{I}_q$ e $\mathbf{R}{=}\sigma_e^2\mathbf{I}_{qm}$. Nell'interpretazione classica quindi, $\mathbf{Z}\boldsymbol{u}$ è una fonte di varianza inserita nel modello che influenza nella medesima maniera un certo insieme di osservazioni, come ad esempio un insieme di studenti provenienti dalla stessa classe o dalla stessa scuola. In questo senso i livelli degli effetti casuali sono estratti da una popolazione ma non sono di interesse in quanto tali ma in quanto derivanti da una popolazione più ampia. Nel tempo l'uso dei modelli ad effetti casuali ha preso un più vasto raggio d'azione in quanto $\mathbf{Z}\boldsymbol{u}$ permette al modello di essere flessibile grazie all'elevato numero di parametri ma allo stesso tempo evita il sovradattamento limitando $\boldsymbol{u}$ tramite la sua covarianza $\mathbf{G}$. La formulazione in (1.2) può essere quindi utilizzata per specificare modelli in cui $\mathbf{Z}\boldsymbol{u}$ non è inteso come un effetto casuale classico (ad esempio le *Splines*).

1.2.1 Stima dei coefficienti degli effetti fissi e effetti casuali

La parte di stima riguarda in particolare gli effetti fissi $\boldsymbol{\beta}$ e la previsione degli effetti casuali $\boldsymbol{u}$ entrambi dipendenti dal parametro, anch'esso da stimare, $\boldsymbol{\phi}{=}(\phi_{\mathbf{R}}, \phi_{\mathbf{G}})$ che

influenza la struttura di varianza rappresentata da $\mathbf{R}$ e $\mathbf{G}$. L'analisi classica procede in tre passi stimando inanzitutto $\boldsymbol{\phi}$ per poi trattarlo come noto e stimare i coefficienti $\boldsymbol{\beta}$ e predire gli effetti casuali $\boldsymbol{u}$. Utilizzando l'impostazione (1.2) si esplicita la densità congiunta (con trasformazione logaritmica) di $\boldsymbol{y}$ e $\boldsymbol{u}$ (entrambe variabili casuali):

$$log\ f(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\phi}) = K - \frac{1}{2}|\mathbf{R}(\phi_{\mathbf{R}})| - \frac{1}{2}|\mathbf{G}(\phi_{\mathbf{G}})| \\ - \frac{1}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{u})^{\intercal}\mathbf{R}(\phi_{\mathbf{R}})^{-1}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{u}) + \boldsymbol{u}^{\intercal}\mathbf{G}(\phi_{\mathbf{G}})^{-1}\boldsymbol{u}. \quad (1.3)$$

Solitamente in una verosimiglianza classica non sono presenti delle variabili casuali non osservate, bensì dei parametri da stimare non noti. L'analisi classica ignora questa stortura trattando la densità in (1.3) come una vera e propria funzione di verosimiglianza per $\boldsymbol{\beta}$ e $\boldsymbol{u}$ con ϕ fissato. Si procede quindi alla stima di $(\boldsymbol{\beta},\boldsymbol{u})$ minimizzando l'opposto della log-verosimiglianza:

$$(\boldsymbol{y} - [\mathbf{X}\ \mathbf{Z}]\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{pmatrix})^{\intercal}\mathbf{R}^{-1}(\boldsymbol{y} - [\mathbf{X}\ \mathbf{Z}]\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{pmatrix}) + (\boldsymbol{\beta}\ \boldsymbol{u})\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{pmatrix}. \quad (1.4)$$

Questa può essere definita come "verosimiglianza penalizzata" in cui il secondo addendo ha la funzione di penalizzare per valori di $\boldsymbol{u}$ grandi rispetto a $\mathbf{G}$. Il concetto che sta alla base della verosimiglianza penalizzata è quella di inserire una parametrizzazione flessibile attraverso un vettore $\boldsymbol{u}$ ad elevata dimensionalità e allo stesso tempo evitare il sovradattamento per mezzo della penalità che costringe $\boldsymbol{u}$ verso lo zero. Derivando (1.4) per $(\boldsymbol{\beta},\boldsymbol{u})$ si ottiene :

$$\frac{\partial log f(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\phi})}{\partial(\boldsymbol{\beta},\boldsymbol{u})} = \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{pmatrix}[\mathbf{X}\ \mathbf{Z}]^{\intercal}\mathbf{R}^{-1}[\mathbf{X}\ \mathbf{Z}] - [\mathbf{X}\ \mathbf{Z}]^{\intercal}\mathbf{R}^{-1}\boldsymbol{y} + \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{pmatrix}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix},$$

e ponendo la derivata uguale a 0 si ottiene

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{u}}\end{pmatrix}_{\boldsymbol{\phi}} = ([\mathbf{X}\ \mathbf{Z}]^{\intercal}\mathbf{R}^{-1}[\mathbf{X}\ \mathbf{Z}] + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix})^{-1}[\mathbf{X}\ \mathbf{Z}]^{\intercal}\mathbf{R}^{-1}\boldsymbol{y}, \quad (1.5)$$

stima che dipende dal parametro $\boldsymbol{\phi}$, considerato in questo passo come noto. Si può notare inoltre che eliminando la penalità si otterrebbe la classica stima ai minimi quadrati generalizzati (MQG), quindi l'aggiunta della penalità va a modificare la stima MQG di $(\boldsymbol{\beta},\boldsymbol{u})$ aggiungendo $\mathbf{G}^{-1}$ alla matrice di precisione di $\boldsymbol{u}$, $\mathbf{Z}^{\intercal}\mathbf{R}^{-1}\mathbf{Z}$ (Hodges, 2014).

1.2.2 Stima di $\boldsymbol{\phi}$ (verosimiglianza vincolata)

L'approccio utilizzato di seguito per la stima di $\boldsymbol{\phi}$ passa attraverso la riformulazione di (1.2) includendo $\mathbf{Z}\boldsymbol{u}$ come parte dell'errore: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$, in cui $\tilde{\boldsymbol{\epsilon}} = \mathbf{Z}\boldsymbol{u} + \epsilon$ ed infine $\mathbf{V}(\boldsymbol{\phi}) = \mathsf{cov}(\tilde{\epsilon}) = \mathbf{Z}\mathbf{G}(\phi_{\mathbf{G}})\mathbf{Z}^{\intercal} + \mathbf{R}(\phi_{\mathbf{R}})$. Questo processo porta alla costruzione della cosidetta "verosimiglianza vincolata" la cui idea fu per la prima volta proposta da Bartlett (1937). La distribuzione di $\boldsymbol{y}$ sarà quindi :

$$\boldsymbol{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\phi})). \tag{1.6}$$

La verosimiglianza sarà in funzione di $(\boldsymbol{\beta}, \boldsymbol{\phi})$ e quindi integrando $\boldsymbol{\beta}$ rimarrà una funzione solamente di $\boldsymbol{\phi}$. La verosimiglianza del modello (1.6) sarà:

$$L(\boldsymbol{\beta}, \mathbf{V}) = \ K|\mathbf{V}|^{\frac{1}{2}} \ \exp(-\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\intercal}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})), \tag{1.7}$$

per integrare $\boldsymbol{\beta}$ è necessario svolgere la forma quadratica e completare il quadrato come segue:

$$\begin{aligned}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\intercal}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \ &= \ \boldsymbol{y}^{\intercal}\mathbf{V}^{-1}\boldsymbol{y} + \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{V}^{-1}\boldsymbol{y} \\ &= \ \boldsymbol{y}^{\intercal}\mathbf{V}^{-1}\mathbf{y} + \widehat{\boldsymbol{\beta}}^{\intercal}\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\intercal}\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}),\end{aligned}$$

in cui $\widehat{\boldsymbol{\beta}} \ = \ (\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{V}^{-1}\boldsymbol{y}$, cioè la stima ai minimi quadrati generalizzati dato $\mathbf{V}$. Integrando $\boldsymbol{\beta}$ e facendo la trasformata logaritmica si ottiene:

$$RL(\boldsymbol{\phi}|\boldsymbol{y}) \propto \frac{1}{2}(log|\mathbf{V}| + log|\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X}| + \boldsymbol{y}^{\intercal}[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}]\boldsymbol{y}. \tag{1.8}$$

Massimizzando $RL(\boldsymbol{\phi}|\boldsymbol{y})$ si ottiene quindi la stima puntuale $\widehat{\boldsymbol{\phi}}$ che può essere utilizzata per il processo di stima visto nel paragrafo precedente al fine di ottenere $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}})$. Il problema principale di questo approccio sta nel fatto che non ci sono alternative al trattare $\boldsymbol{\phi}$ come fosse noto e di conseguenza si ignora l'incertezza in $\phi_{\mathbf{G}}$ e $\phi_{\mathbf{R}}$ e la varianza nelle rispettive stime. Come vedremo nella sezione sucessiva l'approccio bayesiano invece tiene in considerazione questo aspetto (Hodges, 2014).

1.2.3 Interpretazione bayesiana dei modelli a effetti misti

Grazie agli avanzamenti computazionali degli ultimi decenni la statistica bayesiana è molto più utilizzata di un tempo, in questa sezione quindi si descriverà come questa viene utilizzata per la stima dei modelli a effetti misti. In particolare si avrà una funzione

di verosimiglianza $\pi(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \phi_R)$ e le distribuzioni a priori per i parametri non noti $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\phi})$. Una specificazione molto generale presentata in Hodges (2014) è la seguente:

$$\begin{aligned} \pi(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \phi_{\mathbf{R}}) &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u}, \mathbf{R}(\phi_{\mathbf{R}})) \\ \pi(\boldsymbol{u}|\mathbf{G}(\phi_{\mathbf{G}})) &\sim N(0, \mathbf{G}(\phi_{\mathbf{G}})) \\ \pi(\boldsymbol{\beta}) &\sim 1. \end{aligned} \tag{1.9}$$

In questo caso $\pi(\boldsymbol{\beta}) \sim 1$ indica quella che in statistica bayesiana viene definita come una distribuzione a priori non informativa su $\boldsymbol{\beta}$ non limitando in nessun modo la distribuzione dei $\boldsymbol{\beta}$. Per quanto riguarda le distribuzioni a priori per $\phi_{\mathbf{R}}$ e $\phi_{\mathbf{G}}$ sono solitamente indipendenti ma non è chiaro quale sia una priori generalemente efficace, quindi sarà necessario specificarne una ad hoc a seconda del caso. L'obiettivo quindi sarà quello di calcolare la distribuzione a posteriori $\pi(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\phi}|\boldsymbol{y})$, solitamente tramite simulazione. In questo caso a differenza dell'approccio classico i vettori dei parametri non noti $(\boldsymbol{\beta}, \boldsymbol{u})$ sono trattati entrambi come variabili casuali. Le stime puntuali dei parametri sono solitamente calcolate come la media o mediana delle rispettive distribuzioni marginali a posteriori. In questo tipo di approccio la variabilità di $\boldsymbol{\phi}$ viene naturalmente considerata, al contrario di quanto accade nell'approccio classico in cui $\widehat{\boldsymbol{\phi}}$ viene considerato il vero valore del parametro. Inoltre l'analisi non può restituire stime di $\phi_{\mathbf{G}}$ pari a 0 e fornisce degli intervalli sensati dove invece l'approccio classico risulterebbe in una stima uguale a 0. Come accenato in precedenza però il parametro $\boldsymbol{\phi}$ non è privo di complicazioni: nella statistica bayesiana può accadere che la scelta di una certa priori per un parametro incida molto sul risultato finale, ed è questo il caso per $\boldsymbol{\phi}$. Ad oggi non è ancora chiaro quale sia una distribuzione a priori che si comporti in maniera soddisfacente per un ampio ventaglio di situazioni, questo problema è dovuto principalmente al fatto che le strutture di covarianza considerate dai modelli ad effetti misti sono svariate e individuare una priori così duttile risulta proibitivo.

1.2.4 Caso vincolato

Nonostante la formulazione (1.2) sia la più comune in questo contesto è utile introdurre un ulteriore formulazione del modello che in Lee et al. (2006) viene utilizzata come

mezzo per un ampio insieme di analisi. Si definisce il modello in (1.2) con tre equazioni:

$$\begin{aligned} \boldsymbol{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon} & \boldsymbol{\epsilon} &\sim N_{qm}(0, \sigma_e^2 \mathbf{I}_{qm}) \\ \boldsymbol{u} &= \boldsymbol{\delta} & \boldsymbol{\delta} &\sim N_q(0, \sigma_s^2 \mathbf{I}_q) \\ \boldsymbol{\beta} &= \mathbf{M} + \boldsymbol{\xi} & \boldsymbol{\xi} &\sim N_p(0, \sigma_p^2 \mathbf{I}_p). \end{aligned} \tag{1.10}$$

Con una semplice manipolazione matematica si riscrive

$$\mathbf{0}_q = -\boldsymbol{u} + \boldsymbol{\delta}$$

$$M = \boldsymbol{\beta} - \boldsymbol{\xi}.$$

Quindi unendo le precedenti equazioni si ottiene:

$$\left[\begin{array}{c} \boldsymbol{y} \\ \hline \mathbf{0}_q \\ \hline \mathbf{M} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Z} \\ \hline \mathbf{0}_p & -\mathbf{I}_q \\ \hline \mathbf{I}_p & \mathbf{0}_{p\times q} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta} \\ \hline \boldsymbol{u} \end{array}\right] + \left[\begin{array}{c} \boldsymbol{\epsilon} \\ \hline \boldsymbol{\delta} \\ \hline -\boldsymbol{\xi} \end{array}\right], \tag{1.11}$$

con la matrice di varianza e covarianza dei termini di errore:

$$\begin{bmatrix} \mathbf{R} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix}.$$

In questo modo ci si è ricondotti al modello (1.2) con l'aggiunta di una priori $N_p(\mathbf{M}, \boldsymbol{\Sigma})$ su $\boldsymbol{\beta}$. Da qui è intuibile che qualsiasi modello ad effetti misti può essere scritto in questa forma. Questo tipo di formulazione oltre ad essere utile per un più ampio tipo di analisi ci servirà per collegare questa famiglia di modelli con quella dei modelli lineari dinamici (Hodges, 2014).

1.2.5 Modelli lineari dinamici come modelli ad effetti misti

In questa sezione si presentano i modelli lineari dinamici (MLD) e la loro relazione con i modelli ad effetti misti. Questa rappresentazione è utilizzata ad esempio in Rao (1970) per dimostrare dei teoremi Gauss-Markov riguardo i MLD. Questi sono utilizzati per analisi di dati in serie storica e sono strettamente collegati al filtro di Kalman. In particolare questi modelli possono essere usati come filtro per elaborare in tempo reale una stima dello stato del sistema, oppure possono essere utilizzati come lisciatore per

descrivere il comportamento dello stato latente tenendo in considerazione l'intera serie. Si avrà quindi un'equazione che descrive la relazione tra il vettore delle osservazioni al tempo t e il vettore degli stati latenti e una che ne descrive la dinamica evolutiva.

In Hodges (2014) si definisce un MLD per una risposta k-dimensionale $\boldsymbol{y}_t$ come un'equazione per le osservazioni:

$$\boldsymbol{y}_t = \mathbf{F}_t\boldsymbol{\theta}_t + \boldsymbol{n}_t, \qquad \boldsymbol{n_t} \sim N_k(0, \boldsymbol{\Sigma}_t^n), \tag{1.12}$$

in cui $\mathbf{F}_t \in \mathbb{R}^{k\times p}$, $\boldsymbol{\theta}_t \in \mathbb{R}^{p\times 1}$, $\boldsymbol{n}_t \in \mathbb{R}^{k\times 1}$ e $\boldsymbol{\Sigma}_t^n \in \mathbb{R}^{k\times k}$. $\mathbf{F}_t$ è noto, $\boldsymbol{\theta}$ e $\boldsymbol{n}_t$ sono invece non noti ed infine $\boldsymbol{\Sigma}_t^n$ può essere noto, non noto o con solo alcuni elementi noti. L'equazione che invece descrive l'evoluzione dello stato $\boldsymbol{\theta}_t$ è

$$\boldsymbol{\theta}_t = \mathbf{H}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, \qquad \boldsymbol{w}_t \sim N_p(0, \boldsymbol{\Sigma}_t^w), \tag{1.13}$$

anche in questo caso $\mathbf{H}_t$ è una matrice $p \times p$ nota e nuovamente $\boldsymbol{\Sigma}_t^w$ può essere trattata come nota, non nota o con alcuni elementi noti. Solitamente per lo stato iniziale $\boldsymbol{\theta_0}$ viene specificata una distribuzione a priori normale p-variata. Questo modello può essere facilemente riscritto come modello ad effetti misti nella formulazione (1.11).

Si riscrive inanzitutto:

$$\mathbf{0}_p = -\boldsymbol{\theta}_t + \mathbf{H}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t,$$

quindi il MLD può essere ridefinito nel caso vincolato:

$$\left[\begin{array}{c} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_T \\ \hline \mathbf{0}_p \\ \mathbf{0}_p \\ \vdots \\ \mathbf{0}_p \end{array}\right] = \left[\begin{array}{cccccc} \mathbf{0}_{k\times p} & \mathbf{F}_1 & \mathbf{0}_{k\times p} & \dots & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} \\ \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{F}_2 & \dots & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} \\ & \vdots & & \ddots & & \vdots \\ \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \dots & \mathbf{0}_{k\times p} & \mathbf{F}_T \\ \hline \mathbf{H}_1 & -I_p & \mathbf{0}_{k\times p} & \dots & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} \\ \mathbf{0}_{k\times p} & \mathbf{H}_2 & -\mathbf{I}_p & \dots & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} \\ & \vdots & & \ddots & & \vdots \\ \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \dots & \mathbf{H}_T & -\mathbf{I}_p \end{array}\right] \left[\begin{array}{c} \boldsymbol{\theta}_0 \\ \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_T \end{array}\right] + \left[\begin{array}{c} \boldsymbol{n}_1 \\ \vdots \\ \boldsymbol{n}_T \\ \hline \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_T \end{array}\right], \tag{1.14}$$

con gli errori $\boldsymbol{n}_t$ e $\boldsymbol{w}_t$ indipendenti. A questo punto è possibile tornare alla forma ad

effetti misti classica riparametrizzando $\boldsymbol{\theta}_t$ come segue:

$$\begin{aligned}
\boldsymbol{\theta}_1 &= \mathbf{H}_1\boldsymbol{\theta}_0 + \boldsymbol{w}_1 \\
\boldsymbol{\theta}_2 &= \mathbf{H}_2\boldsymbol{\theta}_1 + \boldsymbol{w}_2 \\
&= \mathbf{H}_2\mathbf{H}_1\boldsymbol{\theta}_0 + \mathbf{H}_1\boldsymbol{w}_1 + \boldsymbol{w}_2 \\
&\vdots \\
\boldsymbol{\theta}_t &= \mathbf{H}_t \dots \mathbf{H}_2\mathbf{H}_1\boldsymbol{\theta}_0 + \sum_{i=1}^{t-1}(\mathbf{H}_t \dots \mathbf{H}_{i+1})\boldsymbol{w}_i + \boldsymbol{w}_t.
\end{aligned}$$

Sostituendo quindi nell'equazione (1.12) si ottiene:

$$\begin{aligned}
\boldsymbol{y}_t &= \mathbf{F}_t\boldsymbol{\theta}_t + \boldsymbol{n}_t \\
&= \mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_2\mathbf{H}_1\boldsymbol{\theta}_0 \\
&+ \sum_{i=1}^{t-1} \mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_{i+1}\boldsymbol{w}_i \\
&+ \mathbf{F}_t\boldsymbol{w}_t + \boldsymbol{n}_t.
\end{aligned}$$

In termini di modello a effetti misti quindi avremo $\boldsymbol{\theta}_0$ che rappresenta gli effetti fissi e gli effetti casuali rappresentati invece da $\boldsymbol{w}_t$. Inoltre la matrice di disegno $\mathbf{X}$ degli effetti fissi viene quindi costruita dal prodotto matriciale $\mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_2\mathbf{H}_1$, mentre la matrice di disegno degli effetti casuali $\boldsymbol{Z}$ viene costruita tramite gli altri due addendi come segue:

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_2 & \mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_3 & \mathbf{F}_t\mathbf{H}_t \dots \mathbf{H}_4 & \dots & \mathbf{F}_t\mathbf{H}_t & \mathbf{F}_t \end{bmatrix} \in \mathbb{R}^{k \times pt},$$

che pre-moltiplica il vettore :

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_t \end{bmatrix} \in \mathbb{R}^{pt \times 1}.$$

1.2.6 Esempio: modello di regressione dinamica

Il modello di regressione dinamica per $y_t \in \mathbb{R}$ e $\boldsymbol{x}_t \in \mathbb{R}^p$ estende il modello di regressione statica permettendo ai parametri del modello di seguire una processo stocastico:

$$\begin{aligned} y_t &= \boldsymbol{x}_t^\intercal \boldsymbol{\beta}_t + \sigma_\epsilon \epsilon_t, & \epsilon_t &\sim N(0,1) \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\eta}_t & \boldsymbol{\eta}_t &\sim N(0, \mathbf{I}_p). \end{aligned} \tag{1.15}$$

Il modello in (1.15) può essere riscritto nel caso vincolato come:

$$\left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_T \\ \hline \mathbf{0}_p \\ \mathbf{0}_p \\ \vdots \\ \mathbf{0}_p \end{array}\right] = \left[\begin{array}{ccccc} \boldsymbol{x}_1^\intercal & 0_p & \dots & \dots & 0_p \\ 0_p & \boldsymbol{x}_2^\intercal & \dots & \dots & 0_p \\ \vdots & 0_p & \ddots & \dots & 0_p \\ 0_p & 0_p & 0_p & \dots & \boldsymbol{x}_T^\intercal \\ \hline \mathbf{I}_p & -\mathbf{I}_p & \dots & \dots & 0_{p\times p} \\ 0_{p\times p} & \mathbf{I}_p & -\mathbf{I}_p & \dots & 0_{p\times p} \\ \vdots & \vdots & & \ddots & \vdots \\ 0_{p\times p} & \dots & \dots & \mathbf{I}_p & -\mathbf{I}_p \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_T \end{array}\right] + \left[\begin{array}{c} \epsilon_1 \\ \vdots \\ \epsilon_T \\ \hline \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_T \end{array}\right].$$

Riparametrizziamo nuovamente:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + \sum_{i=1}^{\intercal} \boldsymbol{\eta}_i,$$

e sostituendo:

$$y_t = \boldsymbol{x}_t^\intercal \boldsymbol{\beta}_0 + \boldsymbol{x}_t^\intercal \sum_{i=1}^{\intercal} \boldsymbol{\eta}_i + \sigma_\epsilon \epsilon_t.$$

Quindi $\boldsymbol{\beta}_0$ rappresenta l'effetto fisso e $\boldsymbol{\eta_i}$ l'effetto casuale. Anche in questo caso la distribuzione a priori informativa su $\boldsymbol{\beta_0}$ non è necessaria ma può aiutare la convergenza di algoritmi per l'ottimizzazione della verosimiglianza.

1.3 Stima ai minimi quadrati ricorsivi

Come accennato in precedenza i modelli *state-space* trovano ampio utilizzo nell'analisi delle serie storiche sia per capirne la dinamica sia per scopi previsionali. In generale questi sono un estensione dei classici modelli di regressione e di conseguenza la loro stima condivide delle proprietà con quella ai minimi quadrati. Di seguito si seguirà l'impostazione bayesiana proposta da Triantafyllopoulos (2021). Per collegare i modelli di regressione stimati con il metodo dei minimi quadrati ai modelli *state-space* è utile

considerare i cosidetti minimi quadrati ricorsivi:

$$S(\boldsymbol{\beta}) = \sum_{i=0}^{t-1} \delta^i (y_{t-i} - \boldsymbol{x}_{t-i}^\intercal \boldsymbol{\beta})^2, \tag{1.16}$$

in cui $\delta \in [0,1]$ è un parametro noto che pesa le osservazioni a seconda di quanto si vogliono considerare, nel processo di stima, le osservazioni più lontane nel tempo: tanto più δ sarà vicino a 1 tanto più le osservazioni verranno pesate uniformemente e si otterrà una stima vicina ai classici minimi quadrati. Il modello di regressione sarà quindi :

$$\tilde{\boldsymbol{y}} = \begin{pmatrix} \delta^{(t-1)/2} y_1 \\ \vdots \\ \delta^{1/2} y_{t-1} \\ y_t \end{pmatrix} = \begin{pmatrix} \delta^{(t-1)/2} \boldsymbol{x}_1^\intercal \\ \vdots \\ \delta^{1/2} \boldsymbol{x}_{t-1}^\intercal \\ \boldsymbol{x}_t^\intercal \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.17}$$

differenziando $S(\boldsymbol{\beta})$ si ottiene la stima:

$$\widehat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\intercal \tilde{\boldsymbol{y}} = \left(\textstyle\sum_{i=0}^{t-1} \delta^i \boldsymbol{x}_{t-i} \boldsymbol{x}_{t-i}^\intercal \right)^{-1} \displaystyle\sum_{i=0}^{t-1} \delta^i \boldsymbol{x}_{t-i} y_{t-i}. \tag{1.18}$$

A questo punto è di interesse stimare $\widehat{\boldsymbol{\beta}}_t$ e $\widehat{\sigma}^2$ per ogni tempo, per evitare però di svolgere un inversione di matrice per ogni stima è possibile esprimere le stime al tempo t in funzione di quantità al tempo $t-1$, si definiscono quindi le seguenti quantità:

$$\mathbf{H}_t = \sum_{i=0}^{t-1} \delta^i \boldsymbol{x}_{t-i} \boldsymbol{x}_{t-i}^\intercal = +\boldsymbol{x}_l^\intercal \boldsymbol{x}_t + \delta \mathbf{H}_{t-1} \tag{1.19}$$

$$h_t = \sum_{i=0}^{t-1} \delta^i \boldsymbol{x}_{t-i} y_{t-i} = \boldsymbol{x}_t y_t + \delta h_{t-1} \tag{1.20}$$

$$e_t = y_t - \boldsymbol{x}_t^\intercal \widehat{\boldsymbol{\beta}}_{t-1}. \tag{1.21}$$

Utilizzando quindi le quantità appena specificate si ottengono le stime d'interesse:

- lo stimatore di massima verosimiglianza $\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}}_{t-1} + \mathbf{K}_t e_t$;
- lo stimatore di massima verosimiglianza $\widehat{\sigma}_t^2 = n_t \widehat{\sigma}_{t-1}^2 + r_t e_t$;
- l'aggiornamento ricorsivo di $\mathbf{P}_t = \mathbf{H}_t^{-1} = \frac{1}{\delta}(\mathbf{I} - \frac{\mathbf{P}_{t-1} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{\delta + \boldsymbol{x}_t^\intercal \mathbf{P}_{t-1} \boldsymbol{x}_t}) \mathbf{P}_{t-1}$,

in cui $\mathbf{K}_t = \mathbf{P}_t \boldsymbol{x}_t$, e_t rappresenta l'errore un passo in avanti e $r_t = y_t - \boldsymbol{x}_t^\intercal \widehat{\boldsymbol{\beta}}_t$ l'errore a posteriori, infine in questo iter di stima $\widehat{\boldsymbol{\beta}}_0$,$\mathbf{P}_0$ e $\widehat{\sigma}_0^2$ sono considerati noti. In questo modo

non sono necessarie inversioni di matrici per la stima di $\widehat{\boldsymbol{\beta}}_t$ ed inoltre l'aggiornamento di $\mathbf{P}_t$ evita l'inversione di $\mathbf{H}_t$. Così una volta specificato le quantità assunte note al tempo 0 è possibile avere un algoritmo di stima ricorsiva che inizia stimando $\mathbf{P}_t$, stima $\widehat{\boldsymbol{\beta}}_t$ dopo aver specificato e_t e $\mathbf{K}_t$ e si conclude con la stima di $\widehat{\sigma}_t^2$ per mezzo di n_t e r_t. Si nota come la stima di $\widehat{\boldsymbol{\beta}}_t$ sia un aggiornamento di $\widehat{\boldsymbol{\beta}}_{t-1}$ tramite l'errore a priori e_t (prima di vedere i nuovi dati) pesata per $\mathbf{K}_t$ che rappresenta la media pesata per il fattore δ dei quadrati delle osservazioni $\boldsymbol{x}_t$ fino al tempo t con peso all'osservazione t uguale a 1 e i precedenti di δ^i.

1.4 Modello state-space

A questo punto è possibile notare come il motivo per cui le stime di $\boldsymbol{\beta}$ nella sezione precedente siano dipendenti dal tempo sia l'inserimento del fattore δ che permette un lisciamento locale dei $\boldsymbol{\beta}_t$. Così è possibile interpetare la stima ai minimi quadrati classica come una stima in cui tutti i $\boldsymbol{\beta}_t$ sono uguali.

I modelli *state-space* invece, come visto nel capitolo precedente, ci permettono di definire una dinamica evolutiva dei $\boldsymbol{\beta}_t$ descritta da un processo markoviano, come ad esempio:

$$\begin{aligned} y_t &= \boldsymbol{x}_t^\intercal \boldsymbol{\beta}_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \boldsymbol{\beta}_t &= \mathbf{F}_t^\intercal \boldsymbol{\beta}_{t-1} + \xi_t & \boldsymbol{\xi}_t &\sim N(0, \boldsymbol{\Sigma}_\xi^2), \end{aligned} \tag{1.22}$$

con processo markoviamo si intende che la distribuzione condizionata $\pi(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})$ non dipende dagli stati passati $\boldsymbol{\beta}_{t-2}, \boldsymbol{\beta}_{t-3}, \ldots$. Anche in questo caso $\mathbf{F}_t \in \mathbb{R}^{p\times p}$ è la matrice di transizione e ϵ_t e ξ_t sono dette innovazioni. $\boldsymbol{x}_t \in \mathbb{R}^{p\times 1}$ è invece l'insieme di covariate al tempo t note e $\boldsymbol{\beta}_t \in \mathbb{R}^{p\times 1}$, l'insieme dei coefficienti. Sia la varianza σ_ϵ^2 che la matrice di covarianza $\boldsymbol{\Sigma}_\xi$ possono essere dipendenti dal tempo. Per concludere la specificazione del modello è necessario definire una distribuzione a priori per $\boldsymbol{\beta}_0 \sim N(\widehat{\boldsymbol{\beta}}_{0|0}, \mathbf{P}_{0|0})$, per $\widehat{\boldsymbol{\beta}}_{0|0}$ e $\mathbf{P}_{0|0}$ noti.

1.5 Filtro di Kalman

L'operazione di filtraggio è fondamentale nello studio delle serie storiche in quanto permette di eliminare il rumore in esse presente. Il filtro di Kalman fu per la prima volta formalizzato in Kalman (1960) in un giornale di ingegneria meccanica ed ha due

importanti proprietà: utilizza la forma *state-space* che è in grado di descrivere un elevato numero di fenomeni ed inoltre non ha bisogno dell'assunzione di stazionarietà. Nel proseguo si guarderà al caso in cui le innovazioni seguono una distribuzione normale ma questa assunzione può essere eliminata perdendo così la specificazione della distribuzione condizionata predittiva ma conservando comunque la funzione di filtraggio che hanno i primi due momenti. In sostanza l'algoritmo che segue applica ricorsivimente la distribuzione condizionata a posteriori di $\boldsymbol{\beta}_t$ partendo dalla rispettiva distribuzione di $\boldsymbol{\beta}_{t-1}$, per ogni tempo t partendo da $t = 1$. Considerando quindi il modello (1.22) e (1.23) insieme alla distribuzione a priori di $\boldsymbol{\beta}_0$ si avrà che per ogni tempo:

$$\pi(\boldsymbol{\beta}_t|y_{1:t-1}) \sim N(\widehat{\boldsymbol{\beta}}_{t|t-1}, \mathbf{P}_{t|t-1}) \tag{1.23}$$

$$\pi(\boldsymbol{\beta}_t|y_{1:t}) \sim N(\widehat{\boldsymbol{\beta}}_{t|t}, \mathbf{P}_{t|t}), \tag{1.24}$$

in cui (1.23) indica la distribuzione predittiva di $\boldsymbol{\beta}_t$ al tempo $t-1$ e (1.24) è invece la distribuzione a posteriori di $\boldsymbol{\beta}_t$ al tempo t. Di seguito si indicano le quantità utilizzate per derivare i momenti delle distribuzioni sopra specificate.

$$\widehat{\boldsymbol{\beta}}_{t|t-1} = \mathbf{F}_t\boldsymbol{\beta}_{t-1|t-1} \tag{1.25}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\intercal + \boldsymbol{\Sigma}_\xi^2 \tag{1.26}$$

$$\widehat{\boldsymbol{\beta}}_{t|t} = \widehat{\boldsymbol{\beta}}_{t|t-1} + \mathbf{K}_t e_t \tag{1.27}$$

$$\widehat{y}_{t|t-1} = \boldsymbol{x}_t^\intercal\mathbf{P}_{t|t-1} \tag{1.28}$$

$$e_t = y_t - \widehat{y}_{t|t-1} \tag{1.29}$$

$$q_{t|t-1} = \boldsymbol{x}_t^\intercal\mathbf{P}_{t|t-1}\boldsymbol{x}_t + \sigma_\epsilon^2 \tag{1.30}$$

$$\mathbf{K}_t = \frac{\mathbf{P}_{t|t-1}\boldsymbol{x}_t}{q_{t|t-1}} \tag{1.31}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - q_{t|t-1}\mathbf{K}_t\mathbf{K}_t^\intercal. \tag{1.32}$$

Per la derivazione di queste quantità si rimanda a Triantafyllopoulos (2021). Questo algoritmo quindi partendo dalla specificazione della distribuzione a priori (prima di vedere i dati) di $\boldsymbol{\beta}_0$ deriva $\widehat{\boldsymbol{\beta}}_{1|0}$ e $\mathbf{P}_{1|0}$ e sucessivamente acquisendo la nuova osservazione y_1 calcola $\widehat{\boldsymbol{\beta}}_{1|1}$ e $\mathbf{P}_{1|1}$ le quantità "a posteriori" cioè calcolate tramite l'informazione introdotta dalla nuova osservazione. Il fattore più interessante di questo algoritmo ricorsivo riguarda il fatto che per calcolare $\widehat{\boldsymbol{\beta}}_{t|t}$ e $\mathbf{P}_{t|t}$ si ha bisogno solamente dell'osservazione y_t e delle loro quantità a posteriori del tempo precedente $\widehat{\boldsymbol{\beta}}_{t-1|t-1}$ e $\mathbf{P}_{t-1|t-1}$. È chiaro quindi il parallelo con la stima ai minimi quadrati ricorsivi in quando le stime anche

in questo caso vengono ricavate utilizzando in parte la stima al tempo precedente e in parte quella al tempo t.

1.6 Smoothing

Il Filtro di Kalman oltre a permettere di trasportare $\boldsymbol{\beta}_t$ un tempo avanti utilizzando solamente l'informazione passata, permette anche di stimare $\boldsymbol{\beta}_t$ e y_t utilizzando tutto l'insieme informativo $y_{1:T}$, questo processo viene chiamato *smoothing*. Quello che segue è un algoritmo ricorsivo che procede all'indietro partendo, al tempo T, con $\pi(\boldsymbol{\beta}_T|y_{1:T}) \sim N(\widehat{\boldsymbol{\beta}}_{T|T}|\mathbf{P}_{T|T})$ ottenuto tramite il filtro di Kalman:

$$\pi(\boldsymbol{\beta}_t|y_{1:T}) \sim N(\widehat{\boldsymbol{\beta}}_{t|T}, \mathbf{P}_{t|T}) \tag{1.33}$$

$$\pi(y_t|y_{1:T}) \sim N(\widehat{y}_{t|T}, q_{t|T}), \tag{1.34}$$

in cui (1.33) è la distribuzione lisciata degli stati e (1.34) quella delle osservazioni. Di seguito si definisco le quantità per derivare i momenti delle due distribuzioni:

$$\widehat{\boldsymbol{\beta}}_{t|T} = \widehat{\boldsymbol{\beta}}_{t|t} + \mathbf{L}_t(\widehat{\boldsymbol{\beta}}_{t+1|T} - \widehat{\boldsymbol{\beta}}_{t+1|t}) \tag{1.35}$$

$$\mathbf{P}_{t|T} = \mathbf{P}_{t|t} + \mathbf{L}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{L}_t^\intercal \tag{1.36}$$

$$\mathbf{L}_t = \mathbf{P}_{t|t}\mathbf{F}_{t+1}^\intercal\mathbf{P}_{t+1|t}^{-1} \tag{1.37}$$

$$\widehat{y}_{t|T} = \boldsymbol{x}_t^\intercal\widehat{\boldsymbol{\beta}}_{t|T} \tag{1.38}$$

$$q_{t|T} = \boldsymbol{x}_t^\intercal\mathbf{P}_{t|T}\boldsymbol{x}_t + \sigma_\epsilon^2. \tag{1.39}$$

Il ruolo dello smoothing quindi, data la natura dell'algoritmo, non riguarda la previsione ma può essere utile per la descrizioni di trend o per adattare un modello a dei dati storici. Si può quindi far riferimento alla fase di filtraggio e previsione come una procedura fuori dal campione e alla fase di smoothing come una procedura all'interno del campione, nel primo caso infatti l'obiettivo è stimare l'informazione non ancora acquisita dal modello e nel secondo invece descrive l'informazione già acquisita dal modello (Triantafyllopoulos, 2021).

Capitolo 2

Metodo della direzione alternata dei moltiplicatori

In questo capitolo si presenta in maniera generale un insieme di problemi di minimizzazione utilizzati in statistica e il modo in cui questi vengono estesi in problemi di ottimizzazione vincolata per l'inserimento di informazioni a priori. Inoltre si presenta il metodo della direzione alternata dei moltiplicatori: algoritmo efficiente che permette di stimare modelli di regressione penalizzata.

2.1 Problemi ai minimi quadrati e regolarizzazione

Il recente avanzamento delle tecnologie per la raccolta dei dati hanno permesso negli ultimi anni non solo di conservare grandi moli di dati ma anche di poter raccogliere per ogni osservazione un gran numero di variabili (n e p grandi), questo ha reso centrale l'importanza di modelli capaci di catturare tutta la complessità presente nei dati, ma anche la necessità di sviluppare algoritmi in grado di stimarli su larga scala in maniera efficiente. Molti dei classici problemi di ottimizzazione utilizzati in statistica possono essere ri-espressi come problemi di ottimizzazione vincolata, rendendoli non solo di facile risoluzione ma anche permettendo di aggiungere dell'informazione a priori nel processo di stima (ad esempio la non negatività di un coefficiente) (Byod and Vandenberghe, 2004).

Il più famoso problema di ottimizzazione convessa che sta alla base delle tecniche di regressione è sicuramente il problema ai minimi quadrati, un problema senza vincoli e

con soluzione esplicita.

$$\min_{\boldsymbol{\beta}} \; : \; ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \; = \; \sum_{i=1}^{n} (\boldsymbol{x}_i^\intercal \boldsymbol{\beta} - y_i)^2, \tag{2.1}$$

$\mathbf{X} \in \mathbb{R}^{\,n \times p}$ rappresenta la matrice di disegno e gli $\boldsymbol{x}_i$ le righe di $\mathbf{X}$. Derivando è possibile trovare la soluzione esplicita per $\boldsymbol{\beta}$ che risulta essere $(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{y}$. Fintanto che $n > p$ la soluzione ha senso ed è relativamente semplice da calcolare con un tempo computazionale proporzionale a $p^2 n$. Nonostante sia un metodo molto affidabile e con solide basi teoriche, a volte risulta troppo rigido in applicazioni moderne ad elevata dimensionalità. Un modo comune di rendere il modello più flessibile è quello di aggiungere un termine di regolarizzazione o penalità. Una delle estensioni più comuni è la cosidetta regressione*ridge* per la prima volta proposta in Hoerl and Kennard (2000):

$$\min_{\boldsymbol{\beta}} : \; ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \; + \; \lambda||\boldsymbol{\beta}||_2^2, \tag{2.2}$$

con $\lambda \in \mathbb{R}^+$che aggiunge un termine che penalizza la funzione di costo per valori grandi di $\boldsymbol{\beta}$. Questa formulazione oltre a permettere di risolvere il problema dei minimi quadrati quando la matrice $(\mathbf{X}^\intercal\mathbf{X})$ non è invertibile (sempre nei casi di elevatà dimensionalità), restringe i coefficienti ed evita il sovradattamento della stima. In questo modo quindi si aggiunge dell'informazione a priori riguardo ai coefficienti, nello specifico limitandone l'ampiezza. Essendo in questo caso la penalità differenziabile è facile calcolarne la soluzione esplicita:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \; = \; 2\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\intercal\boldsymbol{y} + 2\lambda\mathbf{I}_p\boldsymbol{\beta},$$

ponendola uguale a 0 si ottiene quindi:

$$\widehat{\boldsymbol{\beta}}_{ridge} \; = \; (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\intercal\boldsymbol{y}, \tag{2.3}$$

in particolare l'aggiunta di $\lambda\mathbf{I}_p$, con $\lambda > 0$, alla diagonale di $\mathbf{X}^\intercal\mathbf{X}$ rende la matrice sempre invertibile. Un'altra penalità utile nei contesti ad alta dimensionalità è la norma l_1:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \; + \; \lambda||\boldsymbol{\beta}||_1, \tag{2.4}$$

questa permette di selezionare $k \leq n < p$ coeffiecienti diversi da 0 con l'obbiettivo di sparsificare la stima, assumendo il fatto che solo una parte delle variabili raccolte durante il campionamento siano utili per spiegare il fenomeno. Questo modello di regressione prende il nome di LASSO (*least absolute shrinkage and selection operator*) formalizzato

per la prima volta da Tibshirani (1996). In questo caso però la penalità rende la funzione non differenziabile ed è quindi utile utilizzare degli algoritmi adatti a problemi più generali per risolverlo.

In generale quindi è possibile esplicitare un problema di minimizzazione in questa forma:

$$\min_{\boldsymbol{\beta}} : ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \ + \sum_{i=1}^{k} \lambda_i \psi_i(\boldsymbol{\beta}), \tag{2.5}$$

in cui $\psi_i(.)$ è una qualsiasi norma applicata ai $\boldsymbol{\beta}$ e $\lambda_i \in \mathbb{R}^+$ coefficiente che pesa la penalità. Si vedrà come sia possibile risolvere questo tipo di problemi riscrivendo il problema in forma vincolata. Riprendendo la sezione precedente è d'interesse notare come i modelli ad effetti misti siano a tutti gli effetti dei modelli ad elevata dimensionalità grazie alla presenza di $\mathbf{Z}\boldsymbol{u}$. In effetti riprendendo la struttura dell'equazione (1.4) è possibile riconoscere la stessa struttura dell'equazione (2.5): con il primo addendo rappresentante il problema ai minimi quadrati ed il secondo la penalità, che in quel caso dipende da $\mathbf{G}$ matrice di covarianza della parte casuale. Inoltre è anche possibile notare come la soluzione per $\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{pmatrix}$ sia molto simile a quella di $\widehat{\boldsymbol{\beta}}_{ridge}$. Questo a riprova di quanto la formalizzazione dei modelli ad effetti misti sia molto generale e flessibile per una vasta varietà di applicazioni.

2.1.1 Problemi vincolati e dual ascent

In seguito si segue la formalizzazione del problema data da Boyd (2010). I problemi di ottimizzazione convessa visti nella sezione precedente possono essere riscritti, in generale, in forma vincolata:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \\ &\textit{sotto il vincolo} : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{y}, \end{aligned} \tag{2.6}$$

con $f : \mathbb{R}^n \to \mathbb{R}$ funzione convessa. Il Langrangiano per il problema (2.6) sarà quindi :

$$L(\boldsymbol{\beta}, \boldsymbol{u}) = f(\boldsymbol{\beta}) + \langle \boldsymbol{u}^\intercal, \mathbf{A}\boldsymbol{\beta} - \boldsymbol{y} \rangle, \tag{2.7}$$

la funzione duale sarà invece:

$$\begin{aligned} g(\boldsymbol{u}) &= \inf_{\boldsymbol{\beta}} \; L(\boldsymbol{\beta}, \boldsymbol{u}) \\ &= \min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) + \langle \boldsymbol{u}^\intercal, \mathbf{A}\boldsymbol{\beta} - \boldsymbol{y} \rangle \\ &= -\max_{\boldsymbol{\beta}} \; (-f(\boldsymbol{\beta}) - \langle \mathbf{A}^\intercal \boldsymbol{u}, \boldsymbol{\beta} \rangle) - \langle \boldsymbol{u}, \boldsymbol{y} \rangle \\ &= -f^*(-\mathbf{A}^\intercal \boldsymbol{u}) - \boldsymbol{y}^\intercal \boldsymbol{u}, \end{aligned} \tag{2.8}$$

con $\boldsymbol{u}$ moltiplicatori di Langrange, f^* è la coniugata convessa di f e $g(\boldsymbol{u})$ è concava e il suo dominio sarà $D = \{\boldsymbol{u} | g(\boldsymbol{u}) > -\infty\}$.

Il problema duale sarà quindi:

$$\max_{\boldsymbol{u}} \; g(\boldsymbol{u}),$$

con $\boldsymbol{u} \in \mathbb{R}^n$. Assumendo ci sia dualità forte, i valori che risolvono il problema primale coincidono con quelli che risolvono il problema duale. È possibile quindi trovare il punto primale ottimo $\boldsymbol{\beta}^*$ dal punto duale ottimo $\boldsymbol{u}^*$ come segue:

$$\boldsymbol{\beta}^* \; = \; \operatorname*{argmin}_{\boldsymbol{\beta}} \; L(\boldsymbol{\beta}, \boldsymbol{u}^*).$$

Nel metodo della discesa duale questo tipo di problema viene risolto attraverso l'ascesa del gradiente. Assumendo che $g()$ sia differenziabile, è possibile trovare $\boldsymbol{\beta}^+ = \operatorname*{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{u})$ che porta a $\nabla g(\boldsymbol{u}) = \mathbf{A}\boldsymbol{\beta}^+ - \boldsymbol{y}$ cioè ai residui del vincolo. Il metodo dell'ascesa duale consiste in due step iterativi:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{\beta}} \; L(\boldsymbol{\beta}, \boldsymbol{u}^*) \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \alpha^k (\mathbf{A}\boldsymbol{\beta}^{k+1} - \boldsymbol{y}), \end{aligned} \tag{2.9}$$

in cui α^k è l'ampiezza del passo verso la minimizzazione, e k rappresenta la k-esima iterazione. Il primo è un passo di minimizzazione per $\boldsymbol{\beta}$ mentre il secondo è un passo per l'aggiornamento della variabile duale. Il vettore $\boldsymbol{u}$ viene anche interpretato come un vettore di "prezzi", e il suo aggiornamento viene descritto come passo per "l'aggiustamento dei prezzi". Questo metodo può essere utilizzato in alcuni casi anche se g non è differenziabile, in questi casi $(\mathbf{A}\boldsymbol{\beta}^{k+1} - \boldsymbol{y})$ sarà il sub-gradiente di $-g$ ma questo rende la convergenza non monotona al contrario del caso precedente. Comunque sia i casi in cui sono presenti le condizioni per usare il metodo dell'ascesa duale del sub-gradiente sono rari e quindi molte volte questo metodo non può essere utilizzato. Il punto di forza che caratterizza questo metodo riguarda il fatto di poter configurarsi in alcuni

casi come un algoritmo decentralizzato: con tale termine si intende un algoritmo che è in grado di raggiungere un obbiettivo globale attraverso diverse dinamiche locali che comunicano tra loro. Si supponga ad esempio di poter suddividere $\boldsymbol{\beta}$ in sotto-vettori, questo renderebbe f separabile portando a:

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{N} f_i(\boldsymbol{\beta}_i), \tag{2.10}$$

in cui $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$ e la variabile $\boldsymbol{\beta}_i \in \mathbb{R}^{n_i}$ sottovettori di $\boldsymbol{\beta}$. Di conseguenza partizionando $\mathbf{A} = [\mathbf{A}_1 \ldots \ \mathbf{A}_N]$ in modo tale che $\mathbf{A}\boldsymbol{\beta} = \sum_{i=1}^{N} \mathbf{A}_i\boldsymbol{\beta}_i$ il Lagrangiano può essere riscritto come

$$L(\boldsymbol{\beta}, \boldsymbol{u}) = \sum_{i=1}^{N} L_i(\boldsymbol{\beta}_i, \boldsymbol{u}) = \sum_{i=1}^{N} (f_i(\boldsymbol{\beta}_i) + \boldsymbol{u}^\intercal \mathbf{A}_i\boldsymbol{\beta}_i) - (1/N)\boldsymbol{u}^\intercal \boldsymbol{y}. \tag{2.11}$$

Quindi la procedura vista precedentemente si divide in N passi differenti per l'aggiornamento di $\boldsymbol{\beta}$ mentre il passo per il vettore dei prezzi resta uguale andando a raggruppare l'informazione proveniente dagli N passi indipendenti:

$$\begin{aligned} \boldsymbol{\beta}_i^{k+1} &= \underset{\boldsymbol{\beta}_i}{\operatorname{argmin}}\ L_i(\boldsymbol{\beta}_i, \boldsymbol{u}^*) \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \alpha^k(\mathbf{A}\boldsymbol{\beta}^{k+1} - \boldsymbol{y}). \end{aligned} \tag{2.12}$$

In questo caso quindi si fa riferimento al metodo come "decomposizione duale", nel caso generale illustrato ci sarà un passo di aggiornamento globale per $\boldsymbol{u}$ che poi verrà distribuito ad ogni $\boldsymbol{\beta}_i$ per aggiornare le stime locali, le quali poi verranno nuovamente raggruppate nell'aggiornamento globale. Questa struttura è sfruttabile attraverso il calcolo parallelo in cui in ogni processore viene calcolato un $\boldsymbol{\beta}_i$.

Per rendere il metodo della ascesa duale più robusto e in particolare per permettere la convergenza dell'algoritmo senza assunzioni quali la convessità stretta e la finitezza di f, è stato sviluppato il metodo del Lagranigiano aumentato, proposto inizialmente da Hestens (1969) e Powell (1969). Si definsce il Lagrangiano aumentato come:

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{u}) = f(\boldsymbol{\beta}) + \boldsymbol{u}^\intercal(\mathbf{A}\boldsymbol{\beta} - \boldsymbol{y}) + \frac{\rho}{2}||\mathbf{A}\boldsymbol{\beta} - \boldsymbol{y}||_2^2, \tag{2.13}$$

in cui $\rho > 0$ è definito come parametro di penalità. Ponendo ρ=0 si ritorna alla forma standard del Lagrangiano (2.7). Inoltre il Lagrangiano aumentato può essere visto come

il Lagrangiano (2.7) associato al problema:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) + \frac{\rho}{2}||\mathbf{A}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 \\ &\textit{sotto il vincolo} : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{y}, \end{aligned} \tag{2.14}$$

questo di fatto è equivalente al problema (2.6) in quanto, dato il vincolo, per ogni valore ammissibile di $\boldsymbol{\beta}$ il termine aggiuntivo sarà 0. La funzione duale $g_\rho(\boldsymbol{u})$ associata sarà quindi: $\inf_{\boldsymbol{\beta}} L_\rho(\boldsymbol{\beta}, \boldsymbol{u})$.

Il beneficio di aggiungere il termine di penalizzazione si sostanzia nell'indebolimento delle condizioni che permettono la differenziabilità di g_ρ. Applicando il metodo dell'ascesa duale si avrà:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{\beta}} \; L_\rho(\boldsymbol{\beta}, \boldsymbol{u}^k) \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \rho(\mathbf{A}\boldsymbol{\beta}^{k+1} - \boldsymbol{y}), \end{aligned} \tag{2.15}$$

conosciuto, appunto, come "metodo dei moltiplicatori". È possibile notare come rispetto al procedimento (2.9) al primo passo viene sostituito il Lagrangiano con il Lagrangiano aumentanto e al posto di α^k si utilizzi invece il termine di penalità ρ. Nonostante questo metodo migliori di molto le proprietà di convergenza dell'algoritmo è presente un lato negativo, infatti quando f è separabile il Lagrangiano aumentano L_ρ non lo è, questo sostanzialmente significa che non è possibile utilizzare la decomposizione duale vista nel paragrafo precedente.

2.2 Metodo delle direzioni alternate dei moltiplicatori

Il metodo delle direzioni alternate dei moltiplicatori (ADMM), proposto per la prima da Glowiski and Marrocco (1975) e Gabay and Mercier (1976), è un algoritmo che unisce la proprietà della decomposizione duale dell'ascesa duale con le proprietà di convergenza del metodo dei moltiplicatori. L'algoritmo risolve problemi nella forma:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + g(\boldsymbol{z}) \\ &\textit{sotto il vincolo} : \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{z} = \boldsymbol{c}, \end{aligned} \tag{2.16}$$

con variabili $\boldsymbol{\beta} \in \mathbb{R}^p$ e $\boldsymbol{z} \in \mathbb{R}^m$, dove $\mathbf{A} \in \mathbb{R}^{n \times p}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ e $\boldsymbol{c} \in \mathbb{R}^n$. Si assume che f e g siano convesse. L'unica differenza con il problema (2.6) della sezione precedente riguarda il fatto che $\boldsymbol{\beta}$ sia stata divisa in due variabili, $\boldsymbol{\beta}$ e $\boldsymbol{z}$, e di conseguenza questo accade anche al vincolo. Come visto in precedenza si definisce il Lagrangiano aumentato come:

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{u}) = f(\boldsymbol{\beta}) + g(\boldsymbol{z}) + \boldsymbol{u}^\intercal(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c}) + \frac{\rho}{2}||\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c}||_2^2, \tag{2.17}$$

ADMM consisterà quindi:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ L_\rho(\boldsymbol{\beta}, \boldsymbol{z}^k, \boldsymbol{u}^k) \\ \boldsymbol{z}^{k+1} &= \underset{\boldsymbol{z}}{\operatorname{argmin}}\ L_\rho(\boldsymbol{\beta}^{k+1}, \boldsymbol{z}, \boldsymbol{u}^k) \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \rho(\mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}). \end{aligned} \tag{2.18}$$

L'algoritmo, come accenato in precedenza, è molto simile all'ascesa duale e al metodo dei moltiplicatori: consiste in un passo per la minimizzazione di $\boldsymbol{\beta}$, uno per $\boldsymbol{z}$, e infine l'aggiornamento della variabile duale. Il metodo dei moltiplicatori si configurerebbe per il problema (2.5) come:

$$\begin{aligned} (\boldsymbol{\beta}^{k+1}, \boldsymbol{z}^{k+1}) &= \underset{\boldsymbol{\beta}, \boldsymbol{z}}{\operatorname{argmin}} L_\rho(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{u}^k) \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \rho(\mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}), \end{aligned}$$

in questo caso quindi l'aggiornamento delle due variabili primali avviene congiuntamente, contrariamente nell'ADMM, $\boldsymbol{\beta}$ e $\boldsymbol{z}$ sono aggiornate in maniera "alternata". Questa separazione in due passi distinti di aggiornamento per le variabili primali è la caratteristica che permette la decomposizione quando f e g sono separabili.

Si definisce di seguito una forma dell'ADMM lievemente diversa ma che sarà più conveniente in alcuni casi, che aggiunge un termine quadratico e scala le variabili duali. Definendo il redisuo $\boldsymbol{r} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c}$, si avrà:

$$\begin{aligned} \boldsymbol{u}^\intercal\boldsymbol{r} + \frac{\rho}{2}||\boldsymbol{r}||_2^2 &= \frac{\rho}{2}||\boldsymbol{r} + \frac{1}{\rho}\boldsymbol{u}||_2^2 - \frac{1}{2\rho}||\boldsymbol{u}||_2^2 \\ &= \frac{\rho}{2}||\boldsymbol{r} + \tilde{\boldsymbol{u}}||_2^2 - \frac{\rho}{2}||\tilde{\boldsymbol{u}}||_2^2, \end{aligned}$$

in cui $\tilde{\boldsymbol{u}} = \frac{1}{\rho}\boldsymbol{u}$ è possibile riscrivere quindi la procedura iterativa come:

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ (f(\boldsymbol{\beta}) + \frac{\rho}{2}||\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{z}^k - \boldsymbol{c} + \tilde{\boldsymbol{u}}^k||_2^2) \\
\boldsymbol{z}^{k+1} &= \underset{\boldsymbol{z}}{\operatorname{argmin}}\ (f(\boldsymbol{z}) + \frac{\rho}{2}||\mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z} - \boldsymbol{c} + \tilde{\boldsymbol{u}}^k||_2^2) \\
\tilde{\boldsymbol{u}}^{k+1} &= \tilde{\boldsymbol{u}}^k + \mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}.
\end{aligned} \tag{2.19}$$

2.2.1 Convergenza dell'algoritmo ADMM

Si descrive in questa sezione, riprendendo Boyd (2010), uno dei risultati più generali riguardanti la convergenza dell'algoritmo in questione che si applica ai problemi visti in seguito. La prima assunzione necessaria riguarda le funzioni f e g è:

Assunto 1. *Le funzioni* $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ *e* $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ *sono chiuse, proprie e convesse.*

Questa prima assunzione implica che esistono $\boldsymbol{\beta}$ e $\boldsymbol{z}$ non necessariamente uniche che minimizzano il Lagrangiano aumentato, inoltre l'assunzione riguardante il codominio permette ad esse di essere non differenziabili. La seconda assunzione riguarda il problema vincolato:

Assunto 2. *Il Lagrangiano non aumentato* L_0 *ha un punto di sella.*

Cioè esiste $(\boldsymbol{\beta}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$ per il quale:

$$L_0(\boldsymbol{\beta}^*, \boldsymbol{z}^*, \boldsymbol{y}) \leq L_0(\boldsymbol{\beta}^*, \boldsymbol{z}^*, \boldsymbol{y}^*) \leq L_0(\boldsymbol{\beta}^*, \boldsymbol{z}, \boldsymbol{y}). \tag{2.20}$$

Questo implica che $(\boldsymbol{\beta}^*, \boldsymbol{z}^*)$ è soluzione di (2.16), quindi che $\mathbf{A}\boldsymbol{\beta}^* + \mathbf{B}\boldsymbol{z}^* = c$, $f(\boldsymbol{\beta}^*) < \infty$ e $g(\boldsymbol{z}^*) < \infty$. Infine è possibile dimostrare, per k che va ad infinito, la convergenza di $\boldsymbol{r}^k$ a 0 e la convergenza del valore della funzione $f(\boldsymbol{\beta}^k) + g(\boldsymbol{z}^k)$ e la variabile duale $\boldsymbol{u}$ ai rispettivi valori ottimi.

2.2.2 Condizioni di ottimalità e criteri per fermare l'algoritmo

Le condizioni necessarie e sufficienti di ottimalità per il problema ADMM (2.15) sono l'esistenza primale:

$$\mathbf{A}\boldsymbol{\beta}^* + \mathbf{B}\boldsymbol{z}^* - \boldsymbol{c} = 0, \tag{2.21}$$

e l'esistenza duale:

$$\begin{aligned} 0 &\in \partial f(\boldsymbol{\beta}^*) + \mathbf{A}^\intercal \boldsymbol{u}^* \\ 0 &\in \partial g(\boldsymbol{z}^*) + \mathbf{B}^\intercal \boldsymbol{u}^*. \end{aligned} \tag{2.22}$$

Per definizione $\boldsymbol{z}^{k+1}$ minimizza $L_\rho(\boldsymbol{\beta}^{k+1}, \boldsymbol{z}, \boldsymbol{u}^k)$ quindi si avrà:

$$\begin{aligned} 0 &\in \partial g(\boldsymbol{z}^{k+1}) + \mathbf{B}^\intercal \boldsymbol{u}^k + \rho \mathbf{B}^\intercal (\mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^{k+1} - c) \\ &= \partial g(\boldsymbol{z}^{k+1}) + \mathbf{B}^\intercal \boldsymbol{u}^k + \rho \mathbf{B}^\intercal \boldsymbol{r}^{k+1} \\ &= \partial g(\boldsymbol{z}^{k+1}) + \mathbf{B}^\intercal \boldsymbol{u}^{k+1}. \end{aligned} \tag{2.23}$$

Questo vuol dire che $\boldsymbol{z}^{k+1}$ e $\boldsymbol{u}^{k+1}$ soddisfa sempre (2.21) e di conseguenza anche le due condizioni (2.22). Per definizione $\boldsymbol{\beta}^{k+1}$ minimizza $L_\rho(\boldsymbol{\beta}, \boldsymbol{z}^k, \boldsymbol{u}^k)$, quindi si avrà:

$$\begin{aligned} 0 &\in \partial f(\boldsymbol{\beta}^{k+1}) + \mathbf{A}^\intercal \boldsymbol{u}^k + \rho \mathbf{A}^\intercal (\mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^k - c) \\ &= \partial f(\boldsymbol{\beta}^{k+1}) + \mathbf{A}^\intercal \boldsymbol{u}^k + \rho \boldsymbol{r}^{k+1} + \rho \mathbf{B}(\boldsymbol{z}^k - \boldsymbol{z}^{k+1}) \\ &= \partial f(\boldsymbol{\beta}^{k+1}) + \mathbf{A}^\intercal \boldsymbol{u}^{k+1} + \rho \mathbf{A}^\intercal \mathbf{B}(\boldsymbol{z}^k - \boldsymbol{z}^{k+1}), \end{aligned} \tag{2.24}$$

e con una semplice manipolazione algebrica

$$\rho \mathbf{A}^\intercal \mathbf{B}(\boldsymbol{z}^k - \boldsymbol{z}^{k+1}) \in \partial f(\boldsymbol{\beta}^{k+1}) + \mathbf{A}^\intercal \boldsymbol{u}^{k+1}.$$

Questo significa che la quantità $\boldsymbol{s}^{k+1} = \rho \mathbf{A}^\intercal \mathbf{B}(\boldsymbol{z}^k - \boldsymbol{z}^{k+1})$, detta residuo duale, può essere vista come un residuo per la condizione duale d'esistenza. Inoltre è possibile individuare anche $\boldsymbol{r}^{k+1} = \mathbf{A}\boldsymbol{\beta}^{k+1} + \mathbf{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}$ come residuo primale. Questi residui convergono a 0 all'aumentare delle iterazioni.

Come ogni algoritmo di ottimizzazione è necessario stabilire un criterio per cui terminare il numero di iterazioni. Si dimostra che

$$f(\boldsymbol{\beta}^k) + g(\boldsymbol{z}^k) - p^* \leq -(\boldsymbol{u}^k)^\intercal \boldsymbol{r}^k + (\boldsymbol{\beta}^k - \boldsymbol{\beta}^*)^\intercal \boldsymbol{s}^k, \tag{2.25}$$

in cui p^* indica il valore ottimo della funzione di perdita e $\boldsymbol{\beta}^*$ le variabili corrispondenti. Questa diseguaglianza mostra che quando i residui $\boldsymbol{r}^k$ e $\boldsymbol{s}^k$ sono piccoli anche la differenza tra il valore obiettivo alla k-esima iterazione e il vero valore ottimale dev'essere piccola. D'altra parte però non conoscendo $\boldsymbol{\beta}^*$ non possiamo usare direttamente questa diseguaglianza come criterio di fermata. Se si assume che $(\boldsymbol{\beta}^k - \boldsymbol{\beta}^*)^\intercal$ sia più piccolo di un certo valore d, si ha che

$$f(\boldsymbol{\beta}^k) + g(\boldsymbol{z}^k) - p^* \leq -(\boldsymbol{u}^k)^\intercal \boldsymbol{r}^k + d||\boldsymbol{s}^k||_2 \leq ||\boldsymbol{u}^k||_2 ||\boldsymbol{r}^k||_2 + d||\boldsymbol{s}^k||_2. \tag{2.26}$$

Questo suggerisce quindi che un plausibile criterio di fermata per ADMM sia che i due residui siano abbastanza piccoli:

$$||\boldsymbol{r}^k||_2 \leq \epsilon^{pri} \quad e \quad ||\boldsymbol{s}^k||_2 \leq \epsilon^{dual}, \tag{2.27}$$

dove $\epsilon^{pri} > 0$ e $\epsilon^{dual} > 0$ sono tolleranze per le condizioni di esistenza. Queste possono essere scelte usando rispettivamente un criterio assoluto e uno relativo, come ad esempio:

$$\begin{aligned} \epsilon^{pri} &= \sqrt{n}\epsilon^{ass} + \epsilon^{rel} \ \max\{||\mathbf{A}\boldsymbol{\beta}^k||_2.||\mathbf{B}\boldsymbol{z}^k||_2, ||\boldsymbol{c}||_2\} \\ \epsilon^{dual} &= \sqrt{p}\epsilon^{ass} + \epsilon^{rel} \ ||\mathbf{A}^\intercal \boldsymbol{u}^k||_2. \end{aligned} \tag{2.28}$$

dove $\epsilon^{ass} > 0$ e $\epsilon^{rel} > 0$ (rispettivamente tolleranza assoluta e relativa) sono solitamente fissate dall'utente; in generale queste dipendono dall'applicazione e dalla scala delle variabili.

2.2.3 Esempio: regressione LASSO

Come visto all'inizio del paragrafo (2.1) un modo per assumere che la variabile risposta di interesse sia spiegata solo da un sottoinsieme delle variabili a disposizione è l'inserimento di una penalità l_1 che permette quindi di stimare solo alcuni dei coefficienti $\boldsymbol{\beta}$ diversi da 0. La soluzione a questo problema non è banale, quindi possiamo servici dell'ADMM per risolverlo. Riprendendo (2.4) il problema di ottimizzazione viene scritto come:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{z}||_1 \\ &\textit{sotto il vincolo} : \boldsymbol{\beta} - \boldsymbol{z} = 0. \end{aligned} \tag{2.29}$$

Procediamo quindi ad esplicitare i tre passi necessari per comporre l'ADMM per la stima della regressione LASSO.

Il Lagrangiano aumentato nella forma scalata sarà:

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{z}, \tilde{\boldsymbol{u}}) = \frac{1}{2}||\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \frac{\rho}{2}||\boldsymbol{\beta} - \boldsymbol{z} + \tilde{\boldsymbol{u}}||_2^2 - \frac{\rho}{2}||\tilde{\boldsymbol{u}}||_2^2 + \lambda||\boldsymbol{z}||_1, \tag{2.30}$$

quindi derivando per $\boldsymbol{\beta}$ e ponendo uguale a 0 si avrà:

$$\begin{aligned} \frac{\partial L_\rho(\boldsymbol{\beta}, \boldsymbol{z}, \tilde{\boldsymbol{u}})}{\partial \boldsymbol{\beta}} &= \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\intercal\boldsymbol{y} + \rho\boldsymbol{\beta} + \rho\tilde{\boldsymbol{u}} - \rho\boldsymbol{z} \\ &= \boldsymbol{\beta}(\mathbf{X}^\intercal\mathbf{X} + \rho\boldsymbol{I}) = \mathbf{X}^\intercal\boldsymbol{y} + \rho(\boldsymbol{z} - \tilde{\boldsymbol{u}}) = 0 \end{aligned}$$

ottenendo così:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X} + \rho\boldsymbol{I})^{-1}(\mathbf{X}^\intercal\boldsymbol{y} + \rho(\boldsymbol{z} - \tilde{\boldsymbol{u}})). \tag{2.31}$$

Si può notare quindi che il passo che fa riferimento all'aggiornamento di $\boldsymbol{\beta}$ è sostanzialmente una regressione *ridge* (2.3).

Per quanto riguarda il passo di aggiornamento per $\boldsymbol{z}$ abbiamo una prima parte differenziabile:

$$\begin{aligned} L_\rho(\boldsymbol{z})^1 &= \frac{\rho}{2}||\boldsymbol{\beta} - \boldsymbol{z} + \tilde{\boldsymbol{u}}||_2^2 \\ &= \frac{\rho}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{\beta} + \boldsymbol{z}^\intercal\boldsymbol{z} + \tilde{\boldsymbol{u}}^\intercal\tilde{\boldsymbol{u}} - 2\tilde{\boldsymbol{u}}^\intercal\boldsymbol{z} - 2\boldsymbol{\beta}^\intercal\boldsymbol{z} + 2\boldsymbol{\beta}^\intercal\tilde{\boldsymbol{u}}), \end{aligned}$$

la sua derivata sarà:

$$\frac{\partial L_\rho(\boldsymbol{z})^1}{\partial \boldsymbol{z}} = \rho(\boldsymbol{z} - (\boldsymbol{\beta} + \boldsymbol{u})),$$

la seconda parte inevece $\lambda||\boldsymbol{z}||_1$ non è differenziabile, ma attingendo al concetto di subdifferenziale si ottiene:

$$\frac{\partial \lambda||\boldsymbol{z}||_1}{\partial \boldsymbol{z}} = \begin{cases} \lambda, & se\ \boldsymbol{z} > 0 \\ -\lambda, & se\ \boldsymbol{z} < 0 \\ [-\lambda, \lambda], & se\ \boldsymbol{z} = 0 \end{cases} \tag{2.32}$$

ponendo quindi a zero risulta:

$$0 = \frac{\partial L_\rho(\boldsymbol{z})^1}{\partial \boldsymbol{z}} + \frac{\partial \lambda||\boldsymbol{z}||_1}{\partial \boldsymbol{z}},$$

$$0 = \begin{cases} \rho(\boldsymbol{z} - (\boldsymbol{\beta} + \boldsymbol{u})) - \lambda, & se\ \boldsymbol{z} < 0 \\ [-\rho(\boldsymbol{\beta} + \boldsymbol{u}) - \lambda, -\rho(\boldsymbol{\beta} + \boldsymbol{u}) + \lambda], & se\ \boldsymbol{z} = 0 \\ \rho(\boldsymbol{z} - (\boldsymbol{\beta} + \boldsymbol{u})) + \lambda, & se\ \boldsymbol{z} > 0 \end{cases}$$

Inanzitutto l'intervallo chiuso nel secondo caso sarà un minimo globale:

$$0 \in [-\rho(\boldsymbol{\beta} + \boldsymbol{u}) - \lambda, -\rho(\boldsymbol{\beta} + \boldsymbol{u}) + \lambda]$$

$$-\rho(\boldsymbol{\beta} + \boldsymbol{u}) - \lambda \leq 0$$

$$-\rho(\boldsymbol{\beta} + \boldsymbol{u}) + \lambda \geq 0,$$

si ottiene quindi

$$-\frac{\lambda}{\rho} \leq (\boldsymbol{\beta} + \boldsymbol{u}) \leq \frac{\lambda}{\rho},$$

risolvendo anche per gli altri due casi si avrà:

$$\begin{cases} \widehat{\boldsymbol{z}} = (\boldsymbol{\beta} + \boldsymbol{u}) - \frac{\lambda}{\rho}, & se\ \widehat{\boldsymbol{z}} > 0 \\ \widehat{\boldsymbol{z}} = 0, & se\ -\frac{\lambda}{\rho} \leq (\boldsymbol{\beta} + \boldsymbol{u}) \leq \frac{\lambda}{\rho} \quad ,+ \\ \widehat{\boldsymbol{z}} = (\boldsymbol{\beta} + \boldsymbol{u}) + \frac{\lambda}{\rho}, & se\ \widehat{\boldsymbol{z}} < 0 \end{cases}$$

che sostituendo $\widehat{\boldsymbol{z}}$ porta al cosidetto operatore *soft-thresholding*:

$$S_{\frac{\lambda}{\rho}}(\boldsymbol{\beta} + \boldsymbol{u}) = \begin{cases} (\boldsymbol{\beta} + \boldsymbol{u}) - \frac{\lambda}{\rho}, & se\ (\boldsymbol{\beta} + \boldsymbol{u}) > \frac{\lambda}{\rho} \\ 0, & se\ |(\boldsymbol{\beta} + \boldsymbol{u})| \leq \frac{\lambda}{\rho} \\ (\boldsymbol{\beta} + \boldsymbol{u}) + \frac{\lambda}{\rho}, & se\ (\boldsymbol{\beta} + \boldsymbol{u}) < -\frac{\lambda}{\rho}. \end{cases} \tag{2.33}$$

Questa soluzione è utile nella maggior parte delle soluzioni dell'algoritmo ADMM.

Quindi aggiungendo il passo duale, l'ADMM per la soluzione del problema LASSO sarà:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= (\mathbf{X}^\intercal \mathbf{X} + \rho \boldsymbol{I})^{-1} (\mathbf{X}^\intercal \boldsymbol{y} + \rho(\boldsymbol{z}^k - \tilde{\boldsymbol{u}}^k)) \\ \boldsymbol{z}^{k+1} &= S_{\frac{\lambda}{\rho}}(\boldsymbol{\beta}^{k+1} + \tilde{\boldsymbol{u}}^k) \\ \tilde{\boldsymbol{u}}^{k+1} &= \tilde{\boldsymbol{u}}^k + \boldsymbol{\beta}^{k+1} - \boldsymbol{z}^{k+1}. \end{aligned} \tag{2.34}$$

Capitolo 3

Regressione ridge generalizzata e modelli dinamici

In questo capitolo si vedrà una versione molto generale e flessibile della regressione *ridge* per poi estenderla ad un caso di regressione dinamica. Quest'ultima verrà collegata ad un particolare caso di modello *state-space* in questo modo giustificandone l'esistenza. Si andrà poi ad arricchire il modello aggiungendo delle penalità non differenziabile che verranno stimate attraverso l'algoritmo appena presentato nel capitolo (2). Infine si tratterà il problema della regolarizzazione e selezione del modello.

3.1 Regressione ridge generalizzata

Si riprende in questo capitolo il modello di regressione *ridge* in equazione (2.2) per generalizzarlo ed estenderlo ad un caso di regressione dinamica per lo *smoothing* di serie storiche. In particolare, come esposto in van Wieringen (2021), è possibile riscrivere il problema in equazione (2.2) come segue:

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\intercal \mathbf{W} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\Delta} (\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{3.1}$$

in cui $\mathbf{W} \in \mathbb{R}^{n \times n}$ matrice diagonale che pesa le osservazioni e $\boldsymbol{\Delta} \in \mathbb{R}^{p \times p}$ matrice definita positiva e simmetrica che definisce il tipo di penalità che, nel caso (2.2), sarà uguale a $\lambda \mathbf{I}_p$. La possibilità di definire la penalità in modo così generale permette molta flessibilità sia per differenziare la penalità per singolo elemento del vettore $\boldsymbol{\beta}$ sia per creare penalità che tengano conto della correlazione tra i parametri. Il vettore $\boldsymbol{\beta}_0$ invece può essere visto come una conoscenza a priori su cui viene centrata la deriva dei coefficienti causata da $\boldsymbol{\Delta}$. Come in precedenza l'aggiunta della penalità alla forma

quadratica assicura l'esistenza di una soluzione unica che è derivabile similmente alla regressione *ridge* classica:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \;=\; 2\mathbf{X}^\intercal\mathbf{W}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\intercal\mathbf{W}\boldsymbol{y} + 2\boldsymbol{\Delta}\boldsymbol{\beta} + 2\boldsymbol{\Delta}\boldsymbol{\beta}_0 = 0,$$

ottenendo così:

$$\boldsymbol{\beta}_{Gridge}(\boldsymbol{\Delta}) \;=\; (\mathbf{X}^\intercal\mathbf{W}\mathbf{X} + \boldsymbol{\Delta})^{-1}(\mathbf{X}^\intercal\mathbf{W}\boldsymbol{y} + \boldsymbol{\Delta}\boldsymbol{\beta}_0). \tag{3.2}$$

Prima di proporre una regressione dinamica basata su quanto visto in precedenza è necessario introdurre un caso particolare di regressione ridge, chiamata di fusione, introdotta per la prima volta da Goeman (2008):

$$\min_{\boldsymbol{\beta}} : ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=2}^{p} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}||_2^2, \tag{3.3}$$

con la parte di penalità che possiamo definire come:

$$\lambda \sum_{j=2}^{p} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}||_2^2 = \boldsymbol{\beta}^\intercal \lambda \mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta}, \tag{3.4}$$

in cui $\mathbf{F} \in \mathbb{R}^{\,(p-1)\times p}$:

$$F_{ij} = \begin{cases} 1, & j = i+1 \\ -1, & j = i \\ 0, & \text{altrimenti} \end{cases} \tag{3.5}$$

in questo caso quindi $\boldsymbol{\Delta} = \lambda\mathbf{F}^\intercal\mathbf{F}$ che porta quindi i β_j adiacenti ad essere uguali tra loro. Questo tipo di situazione non è molto comune nelle applicazioni statistiche ma può tornare utile quando le colonne della matrice $\mathbf{X}$ sono effettivamente ordinate per qualche motivo dato dal contesto.

3.1.1 Formulazione regressione fused-ridge dinamica

Il contesto in cui si inserisce il modello proposto riguarda il lisciamento di serie storiche in cui :

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)^\intercal \in \mathbb{R}^{T\times 1}$ vettore di variabili osservate al tempo T;
- $\boldsymbol{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{p,t})^\intercal \in \mathbb{R}^{p\times 1}$ vettore di p covariate osservate al tempo t;

si definisce quindi il modello di regressione:

$$\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta} + \epsilon \qquad con \qquad \epsilon \sim N(0, \mathbf{I}_T), \tag{3.6}$$

in cui :

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & \dots & \dots & \vdots \\ 0 & 0 & \ddots & \dots & \dots & \vdots \\ 0 & 0 & \ddots & \ddots & \dots & \vdots \\ & \vdots & & \dots & \mathbf{Z}_{T-1} & \vdots \\ 0 & 0 & \dots & \dots & 0 & \mathbf{Z}_T \end{bmatrix} \in \mathbb{R}^{T \times Tp} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \beta_{3,1} \\ \vdots \\ \beta_{p,1} \\ \vdots \\ \vdots \\ \beta_{1,T} \\ \vdots \\ \beta_{p,T} \end{pmatrix} \in \mathbb{R}^{pT \times 1}, \tag{3.7}$$

in cui $\mathbf{Z}_t = \boldsymbol{x}_t^\intercal \in \mathbb{R}^{1 \times p}$.
In questo modo definendo una penalità simile a quella *fused* nel problema (3.3) si andrà a rendere le stime di parametri adiacienti, in questo caso temporalmente, correlate tra loro spingendo le loro distanze verso 0. Avremo quindi:

$$\mathbf{F} = \begin{bmatrix} -\mathbf{I}_p & \mathbf{I}_p & 0_{p\times p} & \dots & 0_{p\times p} & 0_{p\times p} \\ 0_{p\times p} & -\mathbf{I}_p & \mathbf{I}_p & \dots & 0_{p\times p} & 0_{p\times p} \\ & \vdots & & \ddots & & \vdots \\ 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & \dots & -\mathbf{I}_p & \mathbf{I}_p \end{bmatrix} \in \mathbb{R}^{\,p(T-1)\times Tp}, \tag{3.8}$$

portando in fine al modello:

$$\min_{\boldsymbol{\beta}} : \frac{1}{2}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2}\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta}, \tag{3.9}$$

con soluzione per $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{DRidge} = (\mathbf{Z}^\intercal\mathbf{Z} + \lambda\mathbf{F}^\intercal\mathbf{F})^{-1}\mathbf{Z}^\intercal\boldsymbol{y}. \tag{3.10}$$

Questo metodo di stima permette quindi di stimare un coefficiente per ogni tempo e per ogni variabile, ricostruendo così la dinamica evolutiva dei coefficienti. La struttura imposta da $\mathbf{F}$ (3.8) permette di imprimere, come nel caso *state-space*, una dinamica

markoviana nei coefficienti rendendo i coefficienti al tempo t dipendenti da quelli al tempo $t-1$. È interessante infine notare come la matrice $\mathbf{F}$ presente in (3.9) sia riscontrabile, come struttura, non solo nei modelli *state-space*, ma più in generale nei modelli gerarchici presentati nel primo capitolo, in particolare nella forma vincolata (1.11).

3.1.2 Interpretazione bayesiana

La regressione *ridge* generalizzata, come quella classica, ha una forte relazione con la regressione lineare bayesiana. Questa, come visto in precendenza, assume che i parametri $\boldsymbol{\beta}$ e σ^2 siano delle variabili aleatorie e come tali abbiamo una distribuzione, considerando come sempre $\mathbf{X}$ e $\boldsymbol{y}$ quantità note. Si definiscono quindi le distribuzioni a priori:

$$\pi(\boldsymbol{\beta}|\sigma^2) \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Delta}^{-1}) \tag{3.11}$$

$$\pi(\sigma^2) \sim IG(\alpha_0, \xi_0), \tag{3.12}$$

quindi $\boldsymbol{\Delta}$ può essere interpretata come la matrice di precisione della distribuzione a priori di $\boldsymbol{\beta}$. La distribuzione a posteriori congiunta di $\boldsymbol{\beta}$ e σ^2, ipotizzando una verosimiglianza gaussiana, sarà:

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \boldsymbol{y}) \propto\ & \sigma^{-n} exp\{-\frac{1}{2}\sigma^{-2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})\} \\ & \times\ \sigma^{-p} exp\{-\frac{1}{2}\sigma^{-2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal\boldsymbol{\Delta}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\} \\ & \times\ (\sigma^2)^{-\alpha_0-1} exp\{-\frac{1}{2}\sigma^{-2}\xi_0\},\end{aligned} \tag{3.13}$$

sviluppando le forme quadratiche e mettendo in evidenza la stima $\widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta})$ si ottiene:

$$\begin{aligned}&(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\intercal\boldsymbol{\Delta}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\ &= \boldsymbol{y}^\intercal\boldsymbol{y} - \boldsymbol{\beta}^\intercal(\mathbf{X}^\intercal\boldsymbol{y} + \boldsymbol{\Delta}\boldsymbol{\beta}_0) - (\boldsymbol{y}^\intercal\mathbf{X} + \boldsymbol{\beta}_0^\intercal\boldsymbol{\Delta})\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal\boldsymbol{\Delta}\boldsymbol{\beta} \\ &= \boldsymbol{y}^\intercal\boldsymbol{y} - \boldsymbol{\beta}^\intercal(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})^{-1}(\mathbf{X}^\intercal\boldsymbol{y} + \boldsymbol{\Delta}\boldsymbol{\beta}_0) - \\ &\quad (\boldsymbol{y}^\intercal\mathbf{X} + \boldsymbol{\beta}_0^\intercal\boldsymbol{\Delta})(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})^{-1}(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal\boldsymbol{\Delta}\boldsymbol{\beta} \\ &= \boldsymbol{y}^\intercal\boldsymbol{y} - \boldsymbol{y}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})^{-1}\mathbf{X}^\intercal\boldsymbol{y} + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta}))^\intercal(\mathbf{X}^\intercal\mathbf{X}+\boldsymbol{\Delta})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta})),\end{aligned}$$

risultato usato per fattorizzare la distribuzione a posteriori come:

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \mathbf{X}) \propto \pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \mathbf{X})\pi(\sigma^2|\boldsymbol{y}, \mathbf{X}),$$

in cui :

$$\pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \mathbf{X}) \propto \exp\{-\frac{1}{2}\sigma^{-2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta}))^\intercal(\mathbf{X}^\intercal\mathbf{X} + \boldsymbol{\Delta})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta}))\}. \tag{3.14}$$

Questo significa che $E(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \mathbf{X}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{\Delta})$.

In questo modo la regressione ridge generalizzata può essere vista come la stima della media a posteriori di $\boldsymbol{\beta}$ quando si specifica una priori gaussiana multivariata sui coefficienti di regressione. Questa visione bayesiana del problema permette di vedere come la scelta degl'iperparametri $\boldsymbol{\beta}_0$ e $\boldsymbol{\Delta}$ inluenzino la forma della distribuzione a posteriori in termini di variabilità e distorsione e quindi l'accuretazza della stima (van Wieringen, 2021).

3.1.3 Connessione con il modello state-space

Questa specificazione della stima *ridge* generalizzata ha una connessione con una determinata specificazione del modello *state-space* in equazione (1.22) che giustifica le successive estensioni. Scrivendo infatti le equazioni in (1.22) in forma compatta:

$$\begin{aligned} \boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta} + \epsilon \qquad & \epsilon \sim N(0, \sigma^2\mathbf{I}_T) \\ \mathbf{H}\boldsymbol{\beta} = \boldsymbol{\eta} \qquad & \boldsymbol{\eta} \sim N_{Tp}(0, \mathbf{S}), \end{aligned} \tag{3.15}$$

e con la distribuzioni degli errrori:

$$\begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\eta} \end{pmatrix} \sim N_{T+Tp}\left(0, \begin{pmatrix} \sigma^2\mathbf{I}_T & 0 \\ 0 & \mathbf{I}_{Tp} \otimes \boldsymbol{\Sigma}_\eta \end{pmatrix}\right), \tag{3.16}$$

in cui le quantità $\mathbf{H}$ e $\mathbf{S}$ sono definite come:

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_p & \dots & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \dots & \mathbf{0}_{p\times p} \\ -\mathbf{T}_2 & \mathbf{I}_p & & \dots & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} \\ & \vdots & & \ddots & & \vdots \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \dots & -\mathbf{T}_T & \mathbf{I}_p \end{bmatrix} \in \mathbb{R}^{pT\times p} \quad \mathbf{S} = \begin{pmatrix} \mathbf{P}_{1|1} & 0 \\ 0 & \boldsymbol{I_p} \otimes \boldsymbol{\Sigma}_\eta \end{pmatrix}, \tag{3.17}$$

Fahrmeir and Kaufmann (1991) definiscono così, tramite un cambio di variabile da $\boldsymbol{\eta}$ a $\boldsymbol{\beta}$ in (3.15) , la densità marginale $f(\boldsymbol{\beta}) = N_{Tp}(\boldsymbol{\beta}, \mathbf{K}^{-1})$ in cui:

$$\mathbf{K} = \begin{bmatrix} \mathbf{T}_2^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{T}_2 + \mathbf{P}_{1|1} & -\mathbf{T}_2 \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} & & & \\ -\mathbf{T}_2 \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} & \mathbf{T}_2^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{T}_2 + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} & \mathbf{T}_3 \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} & & \\ \ddots & \ddots & \ddots & & \\ & & -\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{T}_{T-1} & \mathbf{T}_T^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{T}_T + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} & \mathbf{T}_T^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \\ & & & -\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{T}_T & \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \end{bmatrix}, \tag{3.18}$$

che deriva da $\mathbf{K} = \mathbf{H}^\intercal \mathbf{S}^{-1} \mathbf{H}$. A questo punto avendo le distribuzioni $f(\boldsymbol{y}|\boldsymbol{\beta})$ e $f(\boldsymbol{\beta})$ tramite il teorema di Bayes è possibile derivare direttamente la distribuzione a posteriori $f(\boldsymbol{\beta}|\boldsymbol{y})$, in particolare le due distribuzioni saranno coniugate in quanto gaussiane e la distribuzione a posteriori sarà proporzionale a $N_{Tp}(\widehat{\boldsymbol{\beta}}_T, \widehat{\mathbf{P}}_T^{-1})$ con :

$$\widehat{\mathbf{P}}_T = \mathbf{K} + \mathbf{Z}^\intercal (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_\epsilon^{-1}) \mathbf{Z} \tag{3.19}$$

$$\widehat{\boldsymbol{\beta}}_T = \widehat{\mathbf{P}}_T^{-1} \mathbf{Z}^\intercal (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_\epsilon^{-1}) \boldsymbol{y}. \tag{3.20}$$

A questo punto è possibile esplicitare la connessione tra la stima in equazione (3.20) e (3.9). Si può quindi specificare un modello *state-space* come quello in equazione (3.15) per ottenere una stima simile a quella $\boldsymbol{\beta}_{DRidge}$ per il primo momento della distribuzione a posteriori di $\boldsymbol{\beta}$: si impone quindi $\boldsymbol{\beta}_1$=0 e $\mathbf{P}_{1|1}$=0, $\sigma_\epsilon^2 = 1$ per ogni tempo e la matrice di precisione di $\boldsymbol{\eta}_t$ come $\lambda \mathbf{I}_p$, ed infine imponendo una dinamica random walk ai coefficienti $\boldsymbol{\beta}_t$, cioè $\mathbf{T}_t = \mathbf{I}_p$ per ogni t, si ottiene:

$$\begin{aligned} y_t &= \mathbf{Z}_t \boldsymbol{\beta}_t + \epsilon_t & \epsilon_t &\sim N(0,1) \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \boldsymbol{\eta}_t & \boldsymbol{\eta}_t &\sim N_p(0, \frac{1}{\lambda} \mathbf{I}_p), \end{aligned} \tag{3.21}$$

che implica quindi $f(\boldsymbol{y}|\boldsymbol{\beta}) \sim N_T(\mathbf{Z}\boldsymbol{\beta}, I_T)$.

Come prima quindi riscrivendo $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{\eta} \sim N_{p(T-1)}(0, \mathbf{I}_{(T-1)} \otimes \frac{1}{\lambda} \mathbf{I}_p)$ si ottiene la densità marginale per $\boldsymbol{\beta}$. Con $\mathbf{H}$ che sarà uguale a $\mathbf{F}$ specificata in equazione (3.8). In questo caso rispetto a (3.17), $\mathbf{H}$ perde una riga a causa del fatto che $\mathbf{P}_{1|1}$ è uguale a 0. Questo infine comporta che le stime in equazione (3.19) e (3.20) si modificheranno come segue:

$$\tilde{\mathbf{P}}_T = \lambda \mathbf{F}^\intercal \mathbf{F} + \mathbf{Z}^\intercal \mathbf{Z} \tag{3.22}$$

$$\tilde{\boldsymbol{\beta}}_T = \tilde{\mathbf{P}}_T^{-1} \mathbf{Z}^\intercal \boldsymbol{y}, \tag{3.23}$$

con $\mathbf{K}$ che, dato le imposizioni precedenti, diventerà esattamente $\lambda\mathbf{F}^\intercal\mathbf{F}$ nel problema (3.9) e in particolare:

$$\mathbf{F}^\intercal\mathbf{F} = \begin{bmatrix} \mathbf{I}_p & -\mathbf{I}_p & 0_{p\times p} & \dots & 0_{p\times p} & 0_{p\times p} \\ -\mathbf{I}_p & 2\mathbf{I}_p & -\mathbf{I}_p & \dots & 0_{p\times p} & 0_{p\times p} \\ & \ddots & \ddots & \ddots & & \vdots \\ 0_{p\times p} & 0_{p\times p} & -\mathbf{I}_p & 2\mathbf{I}_p & -\mathbf{I}_p & 0_{p\times p} \\ 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & -\mathbf{I}_p & \mathbf{I}_p \end{bmatrix} \in \mathbb{R}^{Tp\times Tp}. \tag{3.24}$$

3.2 Regolarizzazione di un modello penalizzato

In un modello come quello appena proposto oltre a scegliere il tipo di penalità ($\boldsymbol{\Delta}$) e il metodo di stima per i parametri di regressione, è di focale importanza la scelta del parametro λ in quanto indica l'intensità con cui agisce la penalizzazione. Questa può essere scelta in modi differenti ma il più comune riguarda la performance del modello al di fuori dei dati con cui è stato stimato. Questo processo viene definito "selezione del modello" in cui appunto, per una serie di modelli, viene stimata una metrica e questa viene utilizzata per scegliere il migliore. Come si vedrà in seguito ci sono diversi modi per farlo, in questo caso però la forma della matrice di disegno rende inutilizzabili tecniche come la validazione incrociata, in quanto partizionando la matrice $\mathbf{Z}$ il numero di parametri cambia.

In seguito si segue l'impostazione presente in Hastie et al. (2009), in cui si definisce un insieme di dati $\tau = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$ e il rispettivo errore del modello stimato $\widehat{f}$ come:

$$Err_\tau = \mathsf{E}_{\mathbf{X}^0,\boldsymbol{y}^0}[L(\boldsymbol{y}^0, \widehat{f}(\mathbf{X}^0))|\tau], \tag{3.25}$$

in cui $(\mathbf{X}^0, \boldsymbol{y}^0)$ indicano un nuovo insieme di dati, non utilizzati per stimare il modello. In generale l'errore che il modello commette sui dati utilizzati per stimarlo è definito come :

$$\widetilde{Err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \widehat{f}(\boldsymbol{x}_i)), \tag{3.26}$$

questo sarà più piccolo rispetto al vero errore Err_τ in quanto i dati sono stati utilizzati sia per stimare il modello sia per valutarne l'errore. Questa metrica non è quindi direttamente utilizzabile per la selezione del modello in quanto si prediligerebbero modelli che si adattano fortemente ai dati a discapito delle performance previsionali al di fuori

dal campione, cioè al momento di prevedere una nuova y_i da una nuova $\boldsymbol{x}_i$. In sostanza questo sarà uno stimatore ottimistico di (3.25). Per sottolineare la natura ottimistica dello stimatore in (3.26) si considera quello che viene definito come errore all'interno del campione:

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{\boldsymbol{y}^0}[L(y_i^0, \widehat{f}(\boldsymbol{x}_i))|\tau,] \tag{3.27}$$

in cui $\boldsymbol{y}^0$ indica che si osservano n nuove variabili risposta per ognuno delle covariate $\boldsymbol{x}_i$ per $i = \ 1 \dots n$, è così possibile individuare una discrepanza tra $\tilde{Err}$ e Err_{in} in quanto il primo è distorto verso il basso per quanto detto prima. Possiamo quindi stimare in generale per le perdite quadratiche, ma anche per molte altre, il valore atteso di questa discrepanza che può essere definita come "ottimismo":

$$\mathsf{E}[Err_{in} - \widetilde{Err}] = \omega = \frac{2}{N}\sum_{i=1}^{N} \mathsf{cov}(\widehat{y}_i, y_i). \tag{3.28}$$

Si nota così che l'ammontare di "ottimismo" dipende da quanto le osservazioni y_i influenzano le previsioni $\widehat{y}_i$ o in termini più pratici quanto il modello è sovradattato. A questo punto quindi è possibile scrivere la relazione:

$$\mathsf{E}_{\boldsymbol{y}}(Err_{in}) = \mathsf{E}_{\boldsymbol{y}}(\widetilde{Err}) + 2\frac{d}{n}\sigma_\epsilon^2, \tag{3.29}$$

in cui $\sum_{i=1}^{N} \mathsf{cov}(\widehat{y}_i, y_i)$ è stato sostutuito da d, l'effettivo numero di parametri nel modello, assumendo di trovarci in un modello in cui $\widehat{\boldsymbol{y}}$ sia ottenuto attraverso un metodo di stima lineare. In seguito quindi si vedranno dei metodi per stimare ω che, sommata ad una stima di $\widetilde{Err}$, permette di ottenere un buon criterio di selezione per modelli in contesto di regolarizzazione. D'altra parte metodi come il *bootstrap* e la validazione incrociata sono utilizzati per stimare Err_τ. Come accennato prima solitamente l'errore dentro al campione non è di grande interesse ma può essere d'aiuto in sede di regolarizzazione del modello e porta spesso a buoni risultati, questo perchè quello che conta è l'ampiezza relativa dell'errore e non quella assoluta (Hastie et al., 2009).

3.3 Criteri di informazione e errore in-sample

In contesti classici i criteri di informazione vengono utilizzati quando il modello non è determinato a priori ma questo è scelto tra un insieme più ampio. Non è quindi utilizzabile direttamente la stima di massima verosimiglianza, si deve anche tener conto

della diversa complessità (il numero di parametri) di questi, aggiungendo un termine di penalità (Azzalini and Scarpa, 2012).

In generale quindi un criterio di informazione sarà nella forma:

$$IC = -2log(L(\widehat{\theta}) + f(p). \tag{3.30}$$

In contesti più complessi in cui si effetuano selezione di variabili o comunque si spingono alcuni dei parametri verso lo zero, il numero di parametri viene sostutito da un indice di complessità che dipende dal parametro di regolarizzazione.

Si va quindi a definire :

$$\widehat{Err}_{in} = \widehat{\widetilde{Err}} + \widehat{\omega}, \tag{3.31}$$

in cui $\widehat{\omega}$ come detto prima, rappresenta il valore atteso dell'ottimismo proveniente dal modello stimato.

Il primo criterio in esame è il criterio di informazione di Akaike (AIC):

$$AIC(\lambda) = \widehat{\widetilde{Err}}(\lambda) + 2\frac{d(\lambda)}{N}, \tag{3.32}$$

dove λ è il parametro di regolarizzazione, $\tilde{Err}(\lambda)$ e $d(\lambda)$ sono rispettivamente l'errore all'interno del campione dato il parametro di regolarizzazione λ e la complessità del modello. Si selezionerà quindi il modello con AIC più basso e il rispettivo parametro di regolarizzazione ottimo $\widehat{\lambda}$. Questo criterio è stato per la prima volta proposto da Akaike (1973) e si basa sulla minimizzazione della divergenza di Kullback-Leibler che può essere interpretata come misura della divergenza tra la distribuzione che genererà le future osservazioni y_i e quelle generate dal modello stimato. Si dimostra che il criterio di informazione di Akaike è non consistente, ciò significa che al divergere di n la probabilità che AIC selezioni un modello sovraparametrizzato non converge a 0. Come alternativa a AIC, che supera il problema della non consistenza, è stato proposto da Schwarz (1978) il criterio di informazione bayesiano (BIC) :

$$BIC(\lambda) = N(\tilde{Err}(\lambda) + log(N)\frac{d(\lambda)}{N}). \tag{3.33}$$

Questo criterio tende a penalizzare più severamente i modelli complessi rispetto ad AIC e per questo seleziona modelli più parsimoniosi. Coerentemente con il suo nome BIC deriva da un approccio bayesiano alla selezione del modello. In particolare in Hastie et al. (2009) si assume di avere un insieme di modelli tra cui scegliere M_m,m=1...M

ognuno con i corrispondenti parametri θ_m. Si ha quindi una distribuzione a priori dei parametri $\pi(\theta_m|\mathrm{M}_m)$, la distribuzione a posteriori sarà :

$$\pi(\mathrm{M}_m|\mathbf{Z}) \propto \pi(\mathrm{M}_m)\pi(\mathbf{Z}|\mathrm{M}_m), \tag{3.34}$$

in cui $\mathbf{Z}$ rappresenta l'insieme di stima $\{\boldsymbol{x}_i, y_i\}_i^N$. È di interesse nell'inferenza bayesiana il cosidetto fattore di bayes che indica il contributo dei dati all'interno della quota di bayes:

$$\frac{\pi(\mathrm{M}_m|\mathbf{Z})}{\pi(\mathrm{M}_l|\mathbf{Z})} = \frac{\pi(\mathrm{M}_m)}{\pi(\mathrm{M}_l)} \frac{\pi(\mathbf{Z}|\mathrm{M}_m)}{\pi(\mathbf{Z}|\mathrm{M}_l)} \tag{3.35}$$

$$BF(\mathbf{Z}) = \frac{\pi(\mathbf{Z}|\mathrm{M}_m)}{\pi(\mathbf{Z}|\mathrm{M}_l)}, \tag{3.36}$$

dove (3.35) è la quota di bayes per due modelli (m e l) in competizione e (3.36) il fattore di bayes. A questo punto per proseguire la derivazione di BIC è necessario approssimare $\pi(\mathbf{Z}|\mathrm{M}_m)$, per i dettagli consultare Ripley (1996), in modo da ottenere :

$$log(\pi(\mathbf{Z}|\mathrm{M}_m)) = log(\pi(\mathbf{Z}|\widehat{\theta}_m, \mathrm{M}_m)) - \frac{d_m}{2} log(N) + O(1), \tag{3.37}$$

dove $\widehat{\theta}_m$ è la stima di massima verosimiglianza e d_m il numero dei parametri liberi nel modello M_m, definendo la funzione di perdita d'interesse come $-2log(\pi(\mathbf{Z}|\widehat{\theta}_m, \mathrm{M}_m))$ avremo il criterio BIC in equazione (3.31). In sostanza quindi scegliere il modello con il BIC minore è equivalente a scegliere il modello con la più grande probabilità a posteriori. In generale non c'è un criterio preferibile tra AIC e BIC in quanto nonostante quest'ultimo sia consistente per campioni finiti tende a selezionare modelli troppo parsimoniosi data la grossa penalità data alla complessità del modello. Un altro criterio che però risulta essere equivalente al AIC nel caso di regressione lineare gaussiana è il Cp di Mallows, definito come:

$$Cp(\lambda) = \tilde{Err}(\lambda) + 2\frac{d(\lambda)}{N}. \tag{3.38}$$

3.3.1 Stima del numero effettivo di parametri

Come accennato in precedenza in modelli in cui è presente un parametro di regolarizzazione (penalità) non è appropriato utilizzare p come vero numero di parametri in quanto questa famiglia di modelli, restringendo i parametri verso lo zero, portano con

se una minore complessità. Con metodo di stima lineare si intende:

$$\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}, \tag{3.39}$$

in cui $\boldsymbol{y}$ è il vettore delle variabili risposta, $\widehat{\boldsymbol{y}}$ delle previsioni e $\mathbf{S}$ una matrice di lisciamento $n \times n$ che dipende da $\mathbf{X}$ ma non da $\boldsymbol{y}$. Nel classico modello di regressione lineare questa matrice viene solitamente chiamata $\mathbf{P} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$, ma nella stessa forma possono essere scritte ad esempio la regressione *ridge* e le *splines* di lisciamento cubiche. Il numero effettivo di parametri sarà quindi :

$$df(\mathbf{S}) = tr(\mathbf{S}), \tag{3.40}$$

in cui $tr()$ indica l'operatore traccia cioè la somma degli elementi diagonali di $\mathbf{S}$. Un ultimo criterio che si discute prende il nome di cross-validazione generalizzata GCV, studiato inizialmente da Golub et al. (1979) e le cui proprietà di ottimalità asintotiche sono discusse in Li (1986): in generale la validazione incrociata è un metodo che viene utilizzato per la regolarizzazione di modelli, esso divide l'insieme di stima in K partizioni e per K volte stima il modello su K-1 partizioni per poi valutarlo sulla K-esima partizione rimanente, ottenendo così K metriche che poi verranno mediate. L'obiettivo di questo metodo è stimare l'errore fuori dal campione . Formalmente avremo:

$$CV(\widehat{f}, \lambda) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \widehat{f}^{-K(i)}(\boldsymbol{x}_i, \lambda)), \tag{3.41}$$

in cui $K(i)$ indica l'insieme degli indici rappresentanti la i-esima partizioni dell'insieme di stima, in sostanza $\widehat{f}^{-K(i)}$ è il modello stimato senza la i-esima sezione. Se il numero di partizioni è uguale a n questa procedura prende il nome di *leave-one-out*. Prendendo in considerazione quanto detto dei metodi di stima lineari è possibile scrivere:

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - \widehat{f}^{-i}(\boldsymbol{x}_i, \lambda)]^2 = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{y_i - \widehat{f}(\boldsymbol{x}_i, \lambda)}{1 - \mathbf{S}_{ii}}\Big]^2, \tag{3.42}$$

in cui $\mathbf{S}_{ii}$ è l'elemento diagonale i-esimo della matrice di proiezione. Tale semplificazione per i metodi di stima lineari è molto importante in quanto permette di ridurre l'onere computazionale. Infine l'approssimazione della GCV è definita:

$$GACV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{y_i - \widehat{f}(\boldsymbol{x}_i, \lambda)}{1 - tr(\mathbf{S})/N}\Big]^2, \tag{3.43}$$

in cui si media per osservazione la traccia della matrice $\mathbf{S}$, cioè la stima dei gradi di libertà, invece di valutarne ogni singolo elemento. Questo insieme di metodi quindi verrà utilizzato per la stima del $\widehat{\lambda}$ nel modello in equazione (3.10), in quanto non necessitano di modificare la dimensione della matrice di disegno $\mathbf{Z}$ e possono quindi essere stimati in una procedura in-sample. È giusto ricordare che l'efficacia di questi metodi di selezione diminuisce in contesti in cui $p > n$.

3.3.2 Stima dell'effettivo numero di parametri nel caso della regressione ridge generalizzata

Come si è visto per utilizzare i criteri di informazione presentati è necessario stimare i gradi di libertà del modello in questione. Nel modello lineare classico $\widehat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{y} = \mathbf{S}\boldsymbol{y}$ in cui $\mathbf{S}$ è la matrice di proiezione, nella regressione lineare quindi i gradi di libertà utilizzati dal modello sono $tr(\mathbf{S})$, in aggiunta se $\mathbf{X}$ è a rango pieno questa sarà pari a p. Per analogia la matrice di lisciamento $\mathbf{S}(\lambda)$ per la regressione ridge generalizzata sarà $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \boldsymbol{\Delta})^{-1}\mathbf{X}^\intercal$, e di conseguenza

$$\widehat{df}(\lambda) = tr(\mathbf{S}(\lambda)) = tr(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \boldsymbol{\Delta})^{-1}\mathbf{X}^\intercal). \tag{3.44}$$

Per analogia quindi nel caso in equazione (3.9) la stima di d sarà:

$$\widehat{d}_{DRidge} = tr(\mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z} + \lambda\mathbf{F}^\intercal\mathbf{F})^{-1}\mathbf{Z}^\intercal). \tag{3.45}$$

3.4 Applicazione: modello di regressione per la previsione dell'inflazione

In seguito si vedrà un semplice esempio per esplicitare le differenze tra i vari strumenti per la stima dei coefficienti di regressione dinamici nel tempo. In particolare si utilizzano i dati relativi all'inflazione nell'indice dei prezzi al consumo (CPI), e il tasso d'interesse relativo ai titoli di stato statunitensi entrambi con cadenza trimestrale, dal primo trimestre del 1953 al secondo trimestre del 1980. Dalla figura (3.1) si può ipotizzare che la relazione tra i due indici sia positiva. Si mettono a confronto tre modelli:

- modello $state - space$ con dinamica dei coefficienti random-walk ed errori non correlati;
- minimi quadrati ricorsivi (1.16) con δ=0.5 (in modo tale da usare solo le ultime 2 osservazioni per stimare i coefficienti);

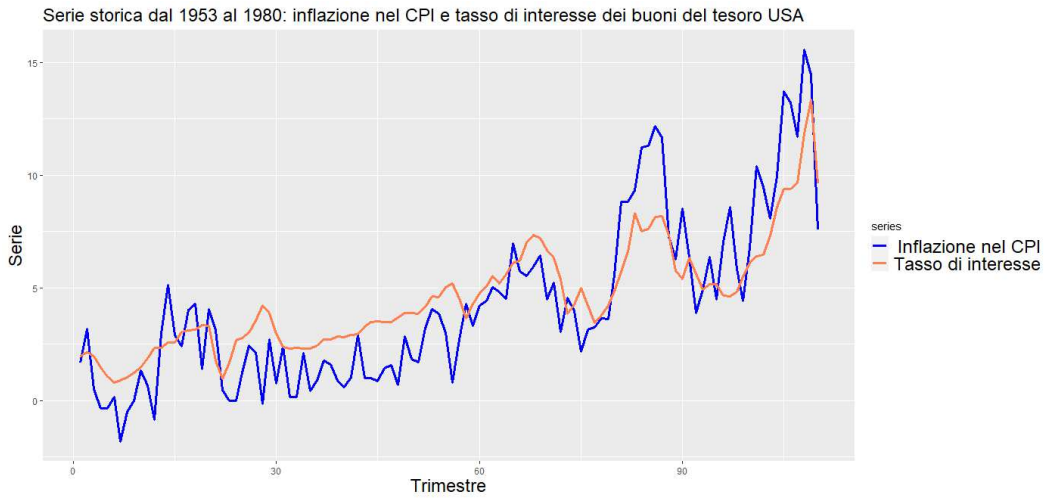

FIGURA 3.1: Serie: inflazione nel CPI e tasso d'interesse dei titoli di stato statunitensi.

- *ridge* generalizzata dinamica (3.9), con $\widehat{\lambda}$ stimato tramite GCV approssimata (3.43).

In particolare per il modello proposto si utilizza una serie di 100 lambda in scala logaritmica da 0.001 e il massimo autovalore di $\mathbf{Z}^\intercal\mathbf{Z}$. Si mostrano di seguito i percorsi dei quattro criteri di informazione esplicitati precedentemente per il caso in questione.

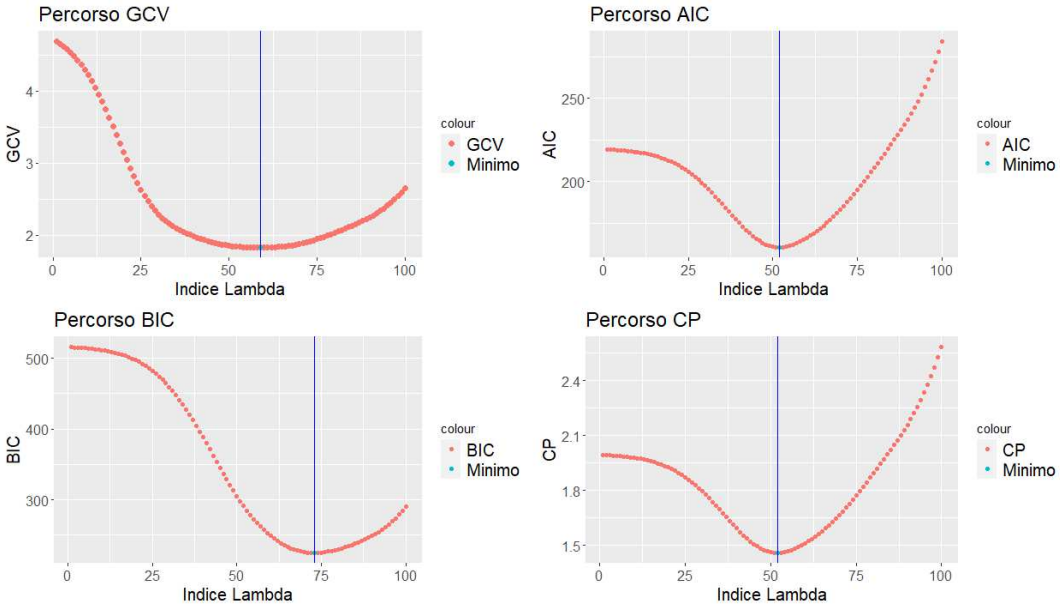

FIGURA 3.2: Percorso Criteri di Informazione rispetto a λ.

Come si può notare dalla figura (3.2) l'unico criterio che si discosta in maniera notevole dagli altri è il BIC: questo infatti è l'indice più severo rispetto al numero di

parametri quindi tende a selezionare modelli più parsimoniosi. Si decide di utilizzare come criterio per la scelta di $\widehat{\lambda}$ il GACV che seleziona un λ=0.46. Per quanto riguarda il modello *state-space* questo viene stimato tramite stima di massima verosimiglianza.

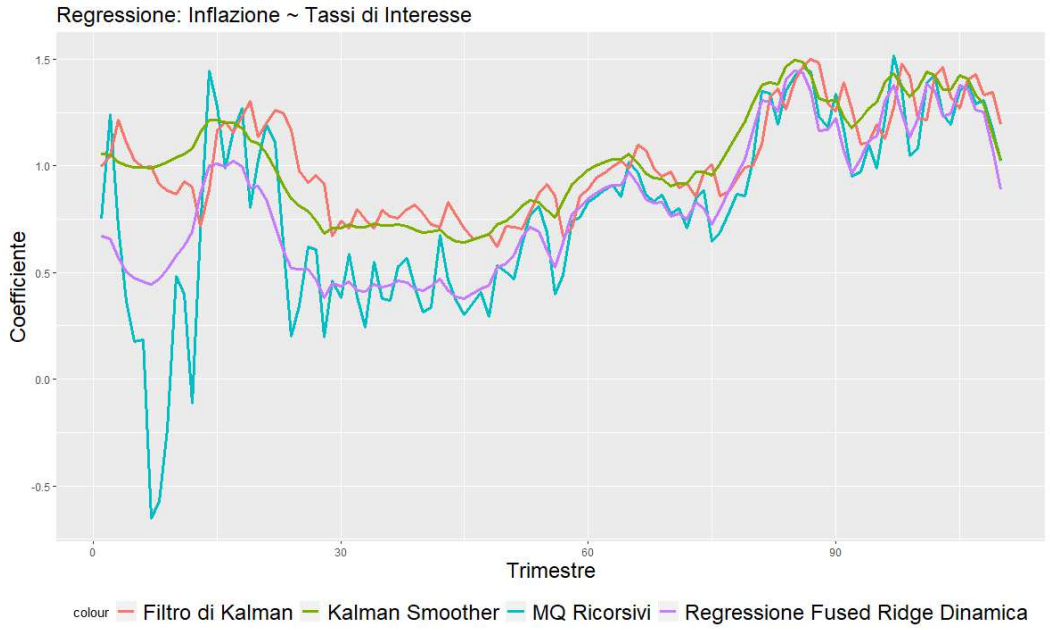

FIGURA 3.3: Dinamica $\widehat{\boldsymbol{\beta}}$: modelli a confronto.

Si può notare dalla figura (3.3) come i quattro metodi divergano sopratutto nella prima metà della serie, il filtro di Kalman e la stima ai minimi quadrati ricorsivi sono quelli con l'andamento più erratico mentre gli altri due, come plausibile, i più lisci. Il modello proposto sembra invece mediare l'andamento dei MQ ricorsivi pur mantendendo un andamento quasi identico allo *smoother* anche se inizialmente traslato verso il basso. Infine è di interesse mostrare quello che in una regressione *ridge* o *LASSO* viene definito come percorso dei coefficienti. In questo caso essendo in un contesto dinamico avremo una serie temporale per λ. Si vede dalla figura (3.4) come all'aumentare della penalità la stima di $\widehat{\boldsymbol{\beta}}$ diventi sempre più liscia.

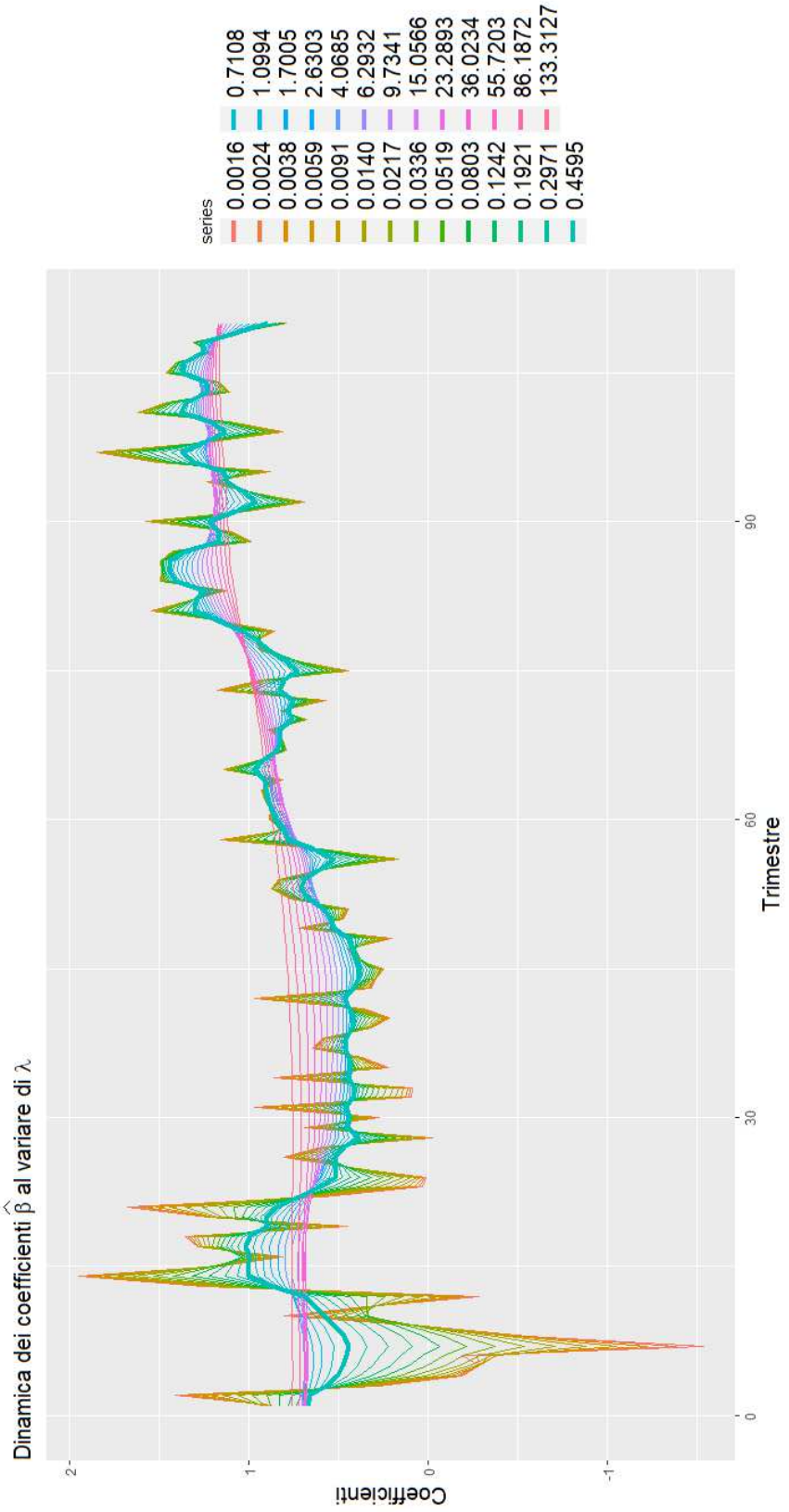

FIGURA 3.4: Percorso coefficienti al variare di λ.

3.5 Confronto tra fused-ridge dinamica e state-space

Di seguito si esegue uno studio di simulazione in cui si indaga l'effetto principalmente di tre parametri: numero di osservazioni/tempo (T), numero di variabili (p) e frazione di parametri uguali a 0 (p_0) sul modello in equazione (3.9). Si prende in considerazione come confronto un modello *state-space* con dinamica degli stati *random-walk* con errori non correlati, stimato con massima verosimiglianza. In particolare si prendono in considerazione la mediana della distanza media in valore assoluto tra i $\boldsymbol{\beta}$ simulati e i $\widehat{\boldsymbol{\beta}}$ stimati dai due modelli che di seguito verrà indicata come *MMAE* in quanto calcolata su tutte le simulazioni riguardanti quella determinata combinazione di parametri. Il motivo per cui si decide di utilizzare la mediana è quello di mitigare l'effetto di selezione del modello errate che portano a volori estremi dell'errore andando ad alterare alcuni dei valori dei *MAE*, di questi comunque verrà tenuto conto nell'analisi della distribuzione dei $\widehat{\lambda}$. In aggiunta verrà anche indicata la deviazione standard (che verrà indicata tra parentesi) delle singole distanze in valore assoluto (*MAE*) per considerare così anche la stabilità di queste. I dati sono simulati come segue:

- viene generata una matrice $\mathbf{X} \in \mathbb{R}^{T \times p}$ in cui ogni riga ($\boldsymbol{x}_t^\intercal$) proviene da $N_p(0_p, \mathbf{I}_p)$;
- vengono generati T errori ϵ_t provenienti da una distribuzione normale standard;
- vengono generati p valori $\boldsymbol{\beta} \in \mathbb{R}^{T \times 1}$ costanti per tutti i tempi provenienti da una $U(-5, 5)$;
- $p_0 \times p$ variabili vengono impostate a 0 per tutta la dinamica;
- i $\boldsymbol{\beta}$ vengono quindi pertubati con shock provenienti da $U(-10, 10)$ in modo da creare due tipi di pattern: a) uno schock presente in mezzo alla dinamica del coefficiente, lasciando l'inizio e la fine costante b) uno shock ad un certo t che proseguirà fino alla fine della dinamica
- una parte dei coefficienti impostati uguali a 0 vengono perturbati come in b) del punto precedente, in modo da aggiungere anche il pattern in cui una coefficiente è uguale a 0 fino ad un tempo t e diverso da 0 per il proseguo ;
- Quindi $y_t = \boldsymbol{x}_t^\intercal \boldsymbol{\beta}_t + \epsilon_t$

In figura (3.5) si mostra un esempio dei coefficienti simulati con p=10, T=200 e $p0$=0.3.

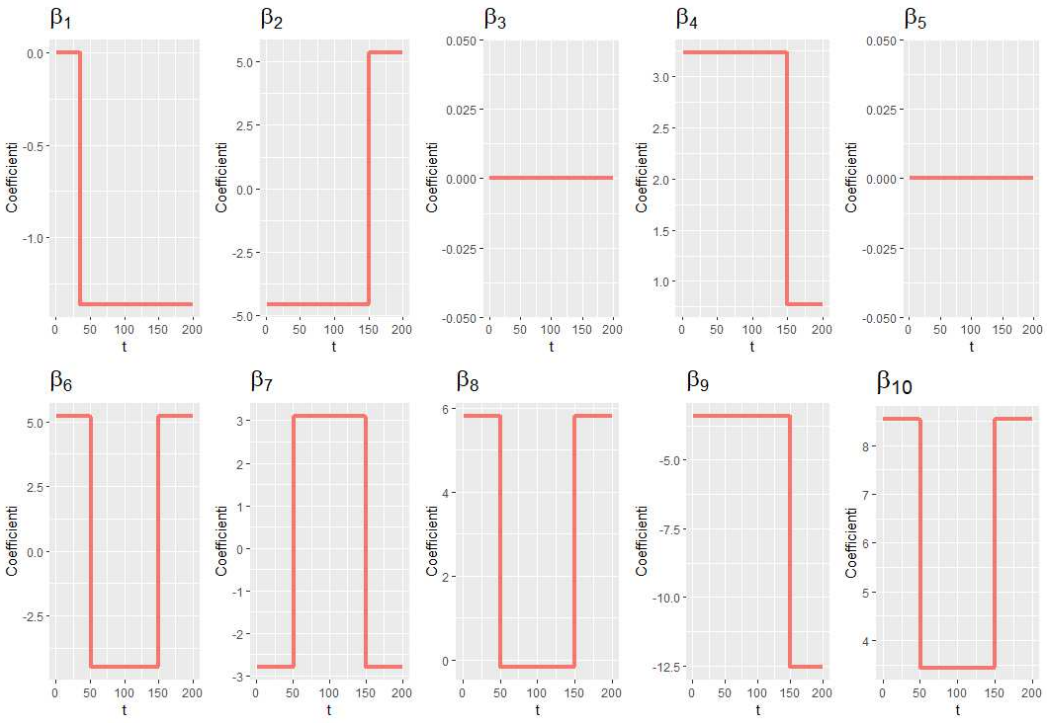

FIGURA 3.5: Esempio Coefficienti simulati T=200, p=10 e p_0=0.3.

Per quanto riguarda la stima del parametri di regolazione $\widehat{\lambda}$ nel modello (3.9) si userà come metrica il GACV (3.43) utilizzando una griglia di 100 valori di λ in scala logaritmica da 0.0001 a 60 (valore determinato empiracamente per permette a tutti i modelli di avere la stessa griglia di lambda per un'analisi successiva), in alternativa può essere impostato come il massimo autovalore di $\mathbf{Z}^\intercal\mathbf{Z}$. In particolare per ogni combinazione di parametri vengono effettuate 50 ripetizioni.

TABELLA 3.1: *Fused-ridge* dinamica

	p_0=0.25			p_0=0.75		
$T\backslash p$	5	15	30	5	15	30
100	0.53 (0.10)	0.99 (0.14)	1.71 (0.20)	0.40 (0.09)	0.73 (0.12)	1.32 (0.20)
200	0.37 (0.08)	0.57 (0.08)	0.94 (0.13)	0.31 (0.07)	0.50 (0.09)	0.76 (0.09)

Tabella 3.2: *Kalman-filter*

	p_0=0.25			p_0=0.75		
$T\backslash p$	5	15	30	5	15	30
100	0.76 (0.19)	1.42 (0.22)	2.1 (0.23)	0.45 (0.09)	0.84 (0.14)	1.37 (0.16)
200	0.53 (0.11)	0.9 (0.13)	1.4 (0.2)	0.33 (0.07)	0.56 (0.1)	0.84 (0.10)

Tabella 3.3: *Kalman-smoother*

	p_0=0.25			p_0=0.75		
$T\backslash p$	5	15	30	5	15	30
100	0.48 (0.11)	0.98 (0.17)	1.75 (0.25)	0.30 (0.07)	0.56 (0.10)	1.05 (0.14)
200	0.33 (0.08)	0.55 (0.09)	0.96 (0.17)	0.22 (0.05)	0.35 (0.07)	0.57 (0.08)

È possibile notare dalle tabelle (3.1), (3.2) e (3.3) che il risultato migliore è raggiunto dal *Kalman-smoother* in termini di MMAE. In tutti i modelli la variazione dei parametri ha l'effetto atteso: l'aumentare del numero di parametri rende le stime più erratiche, l'aumentare invece del numero di tempi diminuisce il MMAE per tutti i modelli. L'effetto di p_0 può sembrare ambiguo in quanto all'aumentare del numero di parametri uguali a 0 diminuisce il MMAE, questo è dovuto semplicemente all'entità in valore assoluto delle stime. Infine uno sguardo alle deviazioni standard dei MMAE calcolati sulle singole ripetizioni mostra una maggiore variabilità per quanto riguarda il *Kalman-filter* rispetto alle altre due stime. È interessante però soffermarsi sul modello proposto, infatti analizzando la distribuzione dei $\widehat{\lambda}$ si nota che per valori di p piccoli (in questo caso 5) nonostante questi siano comunque di piccola entità si distribuiscono in vari valori diversi dal più piccolo (0.0001), ma all'aumentare di p questi iniziano a concentrarsi proprio sul valore più piccolo ad indicare il fatto che il metodo di selezione del modello non funziona come desiderato proprio a causa dell'elevata dimensionalità, le matrici di disegno infatti, ad esempio nel caso di p=15, saranno di dimensioni 100×1500 e 200×3000.

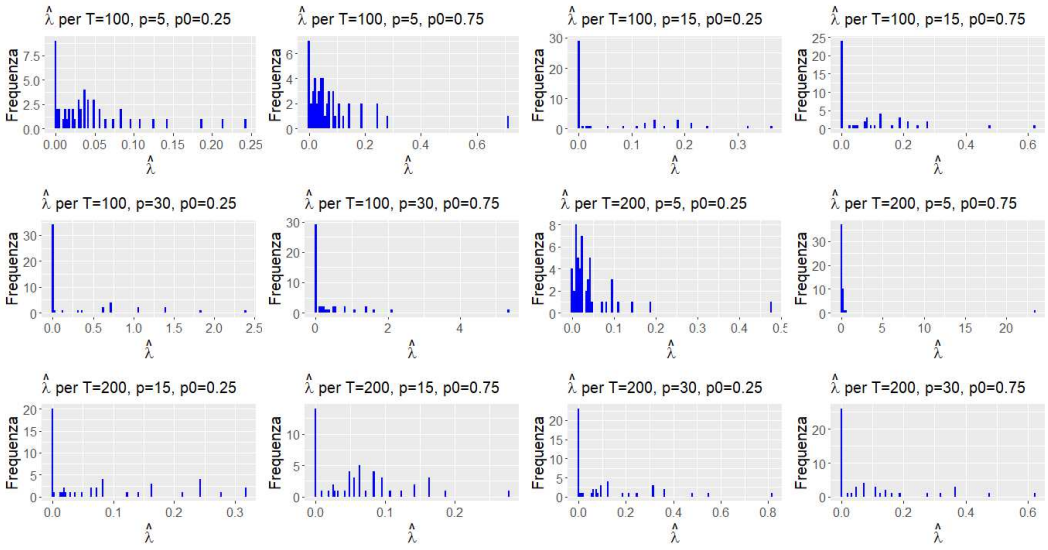

FIGURA 3.6: Distribuzioni parametro di penalità ottimo $\widehat{\lambda}$ al variare dei parametri di simulazione.

Nello specifico osservando la figura (3.6) si nota che i $\widehat{\lambda}$ si concentrano sul valore 0.0001 tanto più p è grande e tanto più T è piccolo sottolineando appunto la vulnerabilità del metodo utilizzato per la selezione dei parametri. È infine interessante notare che l'aumentare della percentuale di coefficienti uguali a 0 fa diminuire la conentrazione di $\widehat{\lambda}$ su 0.0001.

3.6 Fused elastic net dinamica

Seguendo lo sviluppo della penalità detta *elastic-net* presente in Zou and Hastie (2005) si amplia il modello precedente aggiungendo la penalità norma l_1 e di conseguenza un nuovo parametro (α) che pesi i due tipi di penalità:

$$\min_{\boldsymbol{\beta}} : \frac{1}{2}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} + (1-\alpha)\lambda||\mathbf{F}\boldsymbol{\beta}||_1, \tag{3.46}$$

l'aggiunta di questa penalità ha la funzione di rendere alcune delle distanze tra i parametri esattamente zero, il parametro α rende invece le due penalità in competizione andando in sostanza a regolare quanto liscia sarà la stima. Come visto più volte nel capitolo 2 una penalità simile non è differenziabile e quindi è necessario incorrere all'uso di algoritmi efficienti per la soluzione di questo problema di minimizzazione. Si userà quindi l'algoritmo *ADMM* presentato nel capitolo 2. In particolare il problema di ottimizzazione si scriverà:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ \frac{1}{2}||\boldsymbol{y}-\mathbf{Z}\boldsymbol{\beta}||_2^2+\frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta}+(1-\alpha)\lambda||\boldsymbol{z}||_1 \\ &\textit{sotto il vincolo}: \mathbf{F}\boldsymbol{\beta}-\boldsymbol{z}=0, \end{aligned} \tag{3.47}$$

si aggiunge quindi alla forma quadratica la variabile $\boldsymbol{z}$ necessaria per la stima. In questo caso la presenza della matrice $\mathbf{F}$ e del nuovo parametro α non compromette in nessun modo l'algoritmo rispetto al caso *LASSO* visto nella sezione (2.2.3). Una volta scritto il Lagrangiano aumentato come:

$$\begin{aligned} L_\rho(\boldsymbol{\beta},\boldsymbol{z},\tilde{\boldsymbol{u}}) = &\frac{1}{2}||\boldsymbol{y}-\mathbf{Z}\boldsymbol{\beta}||_2^2+\frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} \\ &+\frac{\rho}{2}||\mathbf{F}\boldsymbol{\beta}-\boldsymbol{z}+\tilde{\boldsymbol{u}}||_2^2-\frac{\rho}{2}||\tilde{\boldsymbol{u}}||_2^2+(1-\alpha)\lambda||\boldsymbol{z}||_1, \end{aligned} \tag{3.48}$$

il primo passo dell'algoritmo si deriva similmente a (2.31), ponendo cioè la derivata calcolata rispetto a $\boldsymbol{\beta}$ del Lagrangiano aumentato uguale a 0 e ottenendo:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^\intercal\mathbf{Z}+\gamma\mathbf{F}^\intercal\mathbf{F})^{-1}(\mathbf{Z}^\intercal\boldsymbol{y}+\rho\mathbf{F}(\boldsymbol{z}-\tilde{\boldsymbol{u}})), \tag{3.49}$$

in cui l'unica differenza con il caso *LASSO* riguarda l'utilizzo della matrice $\mathbf{F}$ invece dell'identità e l'aggiunta di $\gamma=(\alpha\lambda+\rho)$. Questo risultato permette una buona flessibilità nella definizione della matrice $\boldsymbol{\Delta}$ anche in presenza della norma l_1. Anche per l'aggiornamento di $\mathbf{Z}$ è sufficiente seguire il processo fatto per il caso *LASSO* con le dovute modifiche. In sostanza, anche in questo caso, utilizzando il sub-gradiente della funzione norma l_1 è possibile derivare la soluzione che si configurerà come (2.33) con la differenza che si avrà $(1-\alpha)\frac{\lambda}{\rho}$ a regolare la soglia e $\mathbf{F}\boldsymbol{\beta}+\boldsymbol{u}$ come argomento. L'algoritmo *ADMM* per il problema (3.47) sarà infine:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= (\mathbf{Z}^\intercal\mathbf{Z}+\gamma\mathbf{F}^\intercal\mathbf{F})^{-1}(\mathbf{Z}^\intercal\boldsymbol{y}+\rho\mathbf{F}^\intercal(\boldsymbol{z}^k-\tilde{\boldsymbol{u}}^k)) \\ \boldsymbol{z}^{k+1} &= S_{\frac{(1-\alpha)\lambda}{\rho}}(\mathbf{F}\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}^k) \\ \tilde{\boldsymbol{u}}^{k+1} &= \tilde{\boldsymbol{u}}^k+\mathbf{F}\boldsymbol{\beta}^{k+1}-\boldsymbol{z}^{k+1}. \end{aligned} \tag{3.50}$$

Questo metodo quindi permette di sorpassare le problematiche che porta con se la norma l_1 senza troppe complicazioni rispetto al caso *LASSO*. A questo punto però nascono due nuove questioni: a) è necessario regolare due parametri invece di uno per individuare il modello migliore b) il metodo di stima non è più lineare e quindi la stima della complessità del modello, necessaria per la costruzione delle metriche usate in precedenza, non sarà più quella vista nel caso ridge generalizzato. In Taylor and Tibishirani (2012) vengono discussi diversi metodi per il calcolo dei gradi di libertà per modelli come

elastic-net, *LASSO* e *generalized LASSO* ma senza coprire il caso proposto in questo elaborato. In particolare la non differenziabilità di $||\mathbf{F}\boldsymbol{\beta}||_1$ pone diversi problemi nella formalizzazione del problema di stima dei gradi libertà.

3.7 MM e stima dell'effettivo numero di parametri

In questa sezione si introdurrà brevemente l'algoritmo "Majorization-Minimization" (MM) per la prima volta proposta in Ortega (1970) per prestarlo al problema di stima della complessità del modello nel caso in cui sia presente la norma l_1 e in particolare nell'applicazione al modello (3.46). Come descritto in Hunter and Lange (2004) l'algoritmo MM può essere utilizzato per la risoluzioni di diversi problemi:

- evitare l'inversioni di matrici ad elevata dimensionalità;
- rendere lineare problemi di ottimizazzione;
- separare i parametri di un problema di ottimizzazione;
- trasformare un problema non differenziabile in uno differenziabile lisciando la funzione.

Utilizzando l'interpretazione in Hunter and Lange (2004) si definisce θ^k come un valore fissato di un parametro non noto θ e $g(\theta|\theta^k)$ una funzione a valori reali di θ la quale forma dipende da θ^k. In questo contesto quindi $g(\theta|\theta^m)$ viene definita come una funzione maggiorante di un'altra funzione a valori reali $f(\theta)$ per un determinato punto θ^k, in sostanza:

$$\begin{aligned} g(\theta|\theta^k) &\geq f(\theta) \qquad \textit{per ogni} \quad \theta, \\ g(\theta^k|\theta^k) &= f(\theta^k). \end{aligned} \tag{3.51}$$

In altre parole quindi la superficie $\theta \to g(\theta|\theta^k)$ giace sopra la superficie $f(\theta)$ ed è tangente ad essa nel punto $\theta = \theta^k$. A questo punto si minimizzerà la funzione $g(\theta|\theta^k)$ invece che quella "originaria" $f(\theta)$. In particolare se si prende in considerazione una funzione $v(\theta)$ due volte differenziabile è possibile maggiorare $v(\theta)$ con una funzione quadratica sufficientemente liscia e tangente in θ^k (Boohning and Lindsay 1988), in termini algebrici se è possibile trovare una matrice $\mathbf{M}$ tale per cui $\mathbf{M}$-$\nabla^2 v(\theta)$ è definita non negativa per ogni θ allora:

$$v(\theta) \leq v(\theta^k) + \nabla v(\theta^k)^\intercal(\theta - \theta^k) + \frac{1}{2}(\theta - \theta^k)^\intercal \mathbf{M}(\theta - \theta^k), \tag{3.52}$$

definisce un limite superiore quadratico. Un contributo quindi di questa tesi è quello di proporre un modo per stimare in maniera approssimata i gradi di libertà di un modello con penalità non differenziabile (in particolare l_1) utilizzando la logica appena descritta riguardante l'algoritmo MM, che quindi verrà utilizzato in questa sede per lisciare una funzione non differenziabile. Un esempio pratico della definizione di un limite superiore quadratico riguarda la funzione $\sqrt{\boldsymbol{\beta}}$ che scrivendo il corrispettivo sviluppo di Taylor si ottiene: $\sqrt{\boldsymbol{\beta}} \leq \boldsymbol{\beta}_0 + (2\sqrt{\boldsymbol{\beta}})^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. A questo punto è possibile procedere con lo sviluppo di Taylor definendo:

$$||\mathbf{F}\boldsymbol{\beta}||_1 \leq ||\mathbf{F}\boldsymbol{\beta}_0||_2 + \frac{||\mathbf{F}\boldsymbol{\beta}||_2^2 - ||\mathbf{F}\boldsymbol{\beta}_0||_2^2}{2||\mathbf{F}\boldsymbol{\beta}_0||_2}, \tag{3.53}$$

in questo modo quindi sostituendo all'interno di (3.46) si ottiene una funzione convessa e differenziabile ed il problema di minimizzazione si riduce ad una serie di norme euclidee:

$$\min_{\boldsymbol{\beta}} : \frac{1}{2}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} + (1-\alpha)\lambda\bigg(||\mathbf{F}\boldsymbol{\beta}_0||_2 + \frac{||\mathbf{F}\boldsymbol{\beta}||_2^2 - ||\mathbf{F}\boldsymbol{\beta}_0||_2^2}{2||\mathbf{F}\boldsymbol{\beta}_0||_2}\bigg), \tag{3.54}$$

è possibile quindi risolvere il problema in equaiozne (3.46) risolvendo (3.54). Nella teoria MM quelli che qui sono indicati come $\boldsymbol{\beta}_0$ farebbero riferimento alla k-esima iterazione, nel nostro caso invece il punto che ci interessa lisciare è quello relativo alla soluzione ottima e quindi $\boldsymbol{\beta}_0$ rappresenterà la soluzione al problema in equazione (3.46). È possibile infine andare a determinare una forma più compatta della funzione (3.55) in modo da esplicitare in maniera più intituitiva la soluzione per la matrice di lisciamento. Si definisce quindi:

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{\beta}_0) &= \frac{1}{2}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} + \widehat{q}_0 + (1-\alpha)\lambda\widehat{q}_1||\mathbf{F}\boldsymbol{\beta}||_2^2 \\ \widehat{q}_0 &= ||\mathbf{F}\boldsymbol{\beta}_0||_2 - \frac{||\mathbf{F}\boldsymbol{\beta}_0||_2^2}{2||\mathbf{F}\boldsymbol{\beta}_0||_2} \qquad \widehat{q}_1 = \frac{1}{2||\mathbf{F}\boldsymbol{\beta}_0||_2}, \end{aligned} \tag{3.55}$$

quindi $\widehat{q}_0$ sarà una costante e non rientrerà nella soluzione per $\boldsymbol{\beta}$. Per individuare quindi la matrice di lisciamento si procede differenziando la funzione in (3.55):

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}} &= \mathbf{Z}^\intercal(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{y}) + \alpha\lambda\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} + 2(1-\alpha)\lambda\widehat{q}_1\mathbf{F}^\intercal\mathbf{F}\boldsymbol{\beta} = 0 \\ \widehat{\boldsymbol{\beta}} &= [\mathbf{Z}^\intercal\mathbf{Z} + \lambda\mathbf{F}^\intercal\mathbf{F}\Big(\alpha + \frac{(1-\alpha)}{||\mathbf{F}\boldsymbol{\beta}_0(\lambda,\alpha)||_2}\Big)\Big]^{-1}\mathbf{Z}^\intercal\boldsymbol{y}, \end{aligned} \tag{3.56}$$

dove è stato esplicitato $\widehat{q_1}$. Per analogia di quanto visto per i modelli descritti in precedenza la stima di d sarà:

$$\widehat{d}(\lambda, \alpha) = tr(\mathbf{Z}\Big[\mathbf{Z}^\intercal\mathbf{Z} + \lambda\mathbf{F}^\intercal\mathbf{F}\Big(\alpha + \frac{(1-\alpha)}{||\mathbf{F}\boldsymbol{\beta}_0(\lambda,\alpha)||_2}\Big)\Big]^{-1}\mathbf{Z}^\intercal). \tag{3.57}$$

3.8 Simulazioni fused elastic net dinamica

Utilizzando quindi gli stessi dati simulati del paragrafo (3.5) si calcolano le stesse metriche utilizzate per i modelli precedenti stimando questa volta il modello in equazione (3.46) e utilizzando come stima di d quella in equazione (3.57).

Tabella 3.4: *Elastic-Fused*

	p_0=0.25			p_0=0.75		
$T\backslash p$	5	15	30	5	15	30
100	0.47 (0.11)	1.13 (0.43)	-	0.32 (0.07)	0.94 (0.43)	-
200	0.31 (0.06)	0.53 (0.08)	-	0.24 (0.05)	0.41 (0.09)	-

Dalla tabella (3.4) si nota come in questo caso il comportamento della stima sia molto più irregolare, infatti per valori di p di 5 e 15 le metriche sembrano migliorare rispetto al modello in equazione (3.9) ma per p pari a 30 la selezione del modello fallisce come accadeva anche nel caso precedente ma in questo caso selezionando il massimo λ possibile invece del più piccolo, selezionando così modelli estremamente penalizzati che portano errori molto ampi. L'aumentare di T mitiga in parte il problema (per p <30) evitando i casi di $\widehat{\lambda}$ vicini a valori estremi. Questo ancora una volta sottolinea la fallibilità di questi metodi di selezione del modello in casi di elevata dimensionalità. Come si vedrà successivamente però l'approssimazione della complessità del modello riguardante il modello in equazione (3.46) stimata tramite la teoria MM sembra raggiungere risultati discreti sopratutto in situazione $p < n$ (si veda capitolo 4).

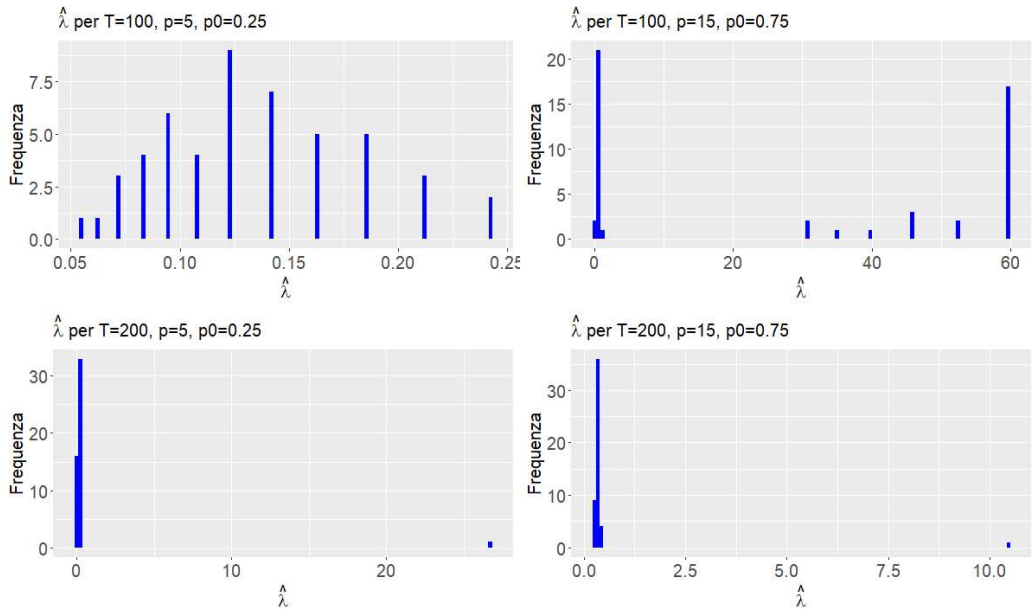

FIGURA 3.7: $\widehat{\lambda}$ al variare dei parametri di simulazione.

Infine per un insieme di dati simulati come in precedenza con T=200 e p=5 si stimano il modello in equazione (3.9) e (3.46) e si mostrano rispettivamente i due migliori modelli in termini di *MAE*. Questi sono stimati sulla stessa griglia di lambda e per quanto riguarda il modello *fused elastic net* fissando $\alpha = 0.7$.

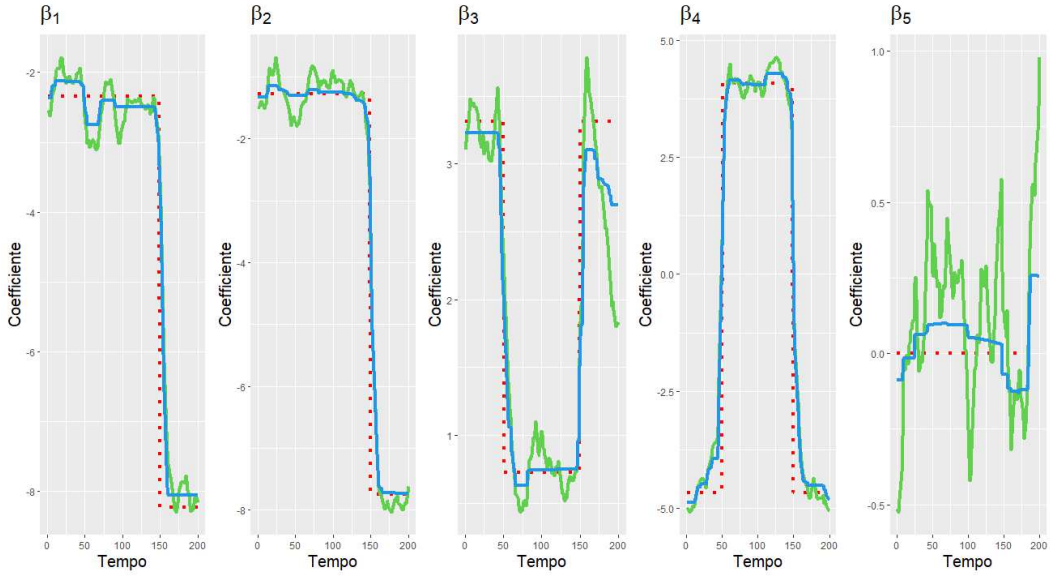

FIGURA 3.8: Confronto migliori modelli su dati simulati, coefficienti simulati (linea rossa tratteggiata), coefficienti modello (3.9) (linea verde), coefficienti modello in equazione (3.46) (linea blu).

Come è possibile notare dalla figura (3.8) le stime riguardanti il modello *fused elastic net* sono molto più liscie e meglio approssimano i $\boldsymbol{\beta}$ simulati, inoltre è interessente vedere come questo aiuti, anche se lievemente, ad inviduare della sparsità nei coefficienti come si può notare da $\boldsymbol{\beta}_5$.

3.9 Group LASSO e generalizzazioni

In questa sezione si descrive il modello formalizzato da Lin (2006) definito come *Group LASSO*, con funzione principalmente interpretativa. Il modello è così definito:

$$\min_{\boldsymbol{\beta}} : \frac{1}{2}\left|\left| \sum_{l=1}^{L} \mathbf{X}_l\boldsymbol{\beta}_l - \boldsymbol{y} \right|\right|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}||\boldsymbol{\beta}_l||_2, \tag{3.58}$$

con l che indica l'appartenenza al gruppo e $\sqrt{p_l}$ che pesa la penalità a seconda dell'ampiezza del gruppo, questa penalità agisce come una penalità *LASSO* ma selezionando gruppi di variabili, invece di singole, che quindi entrano e escono dal modello tutte assieme. Comunque all'interno del gruppo questa penalità non inserisce sparsità, esiste però un estensione del modello , proposta da Friedman et al. (2010), che aggiunge una penalità *LASSO* "pura" che inserisce quindi sparsità anche all'interno dei gruppi mantenendo le caratteristiche sopra citate. Solitamente questo problema viene risolto attraverso l'algoritmo *block coordinate descent* che è un caso particolare dell'algoritmo di norma utilizzato per la stima *LASSO*. È possibile però utilizzare l'algoritmo *ADMM* modificando leggermente la procedura ottenuta in (2.2.3): aggiungendo una matrice $\mathbf{F}_l$ che seleziona le variabili appartenenti al $l-esimo$ gruppo portando così ad avere tanti vincoli quanti sono i gruppi. In pratica si avrà un problema di ottimizzazione vincolata così definito:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \sum_{l=1}^{L} \frac{1}{2}||\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \lambda_l \sum_{l=1}^{L} ||\boldsymbol{z}_l||_1 \\ &\textit{sotto il vincolo} : \mathbf{F}_l\boldsymbol{\beta} - \boldsymbol{z}_l = 0 \qquad l = 1 \dots L, \end{aligned} \tag{3.59}$$

in questo modo il passo di aggiornamento di $\boldsymbol{\beta}$ rispetto al caso classico sostituisce con $\mathbf{F}^\intercal\mathbf{F}$ la matrice identità, in cui $\mathbf{F}$ è la matrice che incolonna le matrici $\mathbf{F}_l$ e il passo di aggiornamento della variabile $\mathbf{Z}$ verrà invece suddiviso in L passi in cui il k-esimo step sarà:

$$\boldsymbol{z}_l^{k+1} = S^*_{\frac{\lambda_l}{\rho}}(\mathbf{F}_l\boldsymbol{\beta}^{k+1} + \tilde{\boldsymbol{u}}_l^k), \tag{3.60}$$

con $\lambda_l = \lambda\sqrt{p_l}$, il passo duale invece resterà invariato raccogliendo l'informazione proveniente dai L passi. Differenza sostanziale nel passo di aggiornamento di $\mathbf{Z}$ sta nell'operatore S che si configura come un operatore *soft-thresholding* a bloccchi definito come:

$$S_k^*(\boldsymbol{a}) = (1 - k/||\boldsymbol{a}||_2)_+ \boldsymbol{a}, \tag{3.61}$$

in cui $S_k^* : \mathbb{R}^m \to \mathbb{R}^m$ che si riduce all'operatore visto nel capitolo 2 se $\boldsymbol{a}$ è scalare. A questo punto il metodo è abbastanza flessibile potendo definire i gruppi di variabili in diversi modi mantenendo fisso il metodo di stima che rimane simile a quello per il *LASSO*, inoltre essendo i passi riguardanti l'aggiornamento di $\mathbf{Z}$ indipendenti è possibile utilizzare un approccio parallelo al calcolo similmente alla procedura descritta in (2.12). Come accennato in precedenza questo metodo ha come principale utilità il fatto di riuscire a scindere l'effetto di insiemi di variabili che vengono raggruppate per tipologia data dal dominio di applicazione, ad esempio in un applicazione economico-sociale i gruppi potrebbero rappresentare variabili reddituali, demografiche, geografiche e così via, in questo modo l'inclusione di un gruppo permette a tutta la tipologia di essere inclusa nel modello dando così spunti interpretativi.

3.10 LASSO adattivo

La stima classica dei modelli lineari ha la caratteristica di essere non distorta, le stime derivanti dai modelli penalizzati visti fin'ora sono invece distorte portando però con sè minore varianza avendo appunto una struttura più flessibile ed un processo di stima che evita il sovradattamento. La distorsione presente in *LASSO* e *ridge* è principalmente dovuta al fatto che i coefficienti vengono ristretti verso lo 0 in maniera lineare con il crescere delle stime di $\boldsymbol{\beta}$, quindi più grande è il coefficiente più grande sarà la penalizzazione. Si sono così sviluppate nel tempo diversi tipi di penalità che agiscono in modo adattivo rispetto alla stima andando a penalizzare di più coefficienti piccoli e di meno coefficienti grandi riducendo così la distorsione, inoltre l'inserimento di questo tipo di penalità permette di distanziare, durante il percorso di regolarizzazione di λ, l'entrata dei coefficienti nel modello assicurando una maggiore precisione nell'individuare il λ ottimo. Il modo più semplice di fare questo è stato proposto da Zou (2012) che sfruttando il fatto che la distorsione dello stimatore dipende essenzialmente da λ

esplicita un modello in cui questo è diverso per ogni $\boldsymbol{\beta}_j$. In particolare si specifica:

$$\min_{\boldsymbol{\beta}} \ \frac{1}{2}||\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \tag{3.62}$$

dove ω_j è un peso che dev'essere tanto più piccolo tanto più grande è β_j andando così a definire λ_j=$\lambda\omega_j$. Non è possibile conoscere a priori i valori o l'ampiezza dei veri valori dei parametri β_j, ma è comunque possibile definire in maniera consistente i pesi ω_j utilizzando una stima a due passi: si usa uno stimatore non distorto di β_j, come ad esempio la stima ai minimi quadrati, e si definisce ω_j come $\frac{1}{|\beta_j^{MQ}|}$, in questo modo affidandosi alle proprietà teoriche della stima MQ si vanno a definire dei pesi che sono più grandi per valori piccoli di β_j^{MQ} e viceversa. In casi in cui però si si trova in constesti di alta dimensionalità in cui non è possibile computare stime non distorte bisogna ricorrere a soluziono subottimali come la stima *ridge*, in ogni caso cè da ricordare l'onere computazionale e il rischio di sovradattamento in cui si incorre nel processo di stima in due passi.

3.11 Fused elastic net con group LASSO per variabile

Si introduce in questa sezione un ampliamento al modello (3.46) con l'obiettivo di inserire sparsità nella stima dei coefficienti dinamici, similmente a quello che accade nel caso *LASSO* statico. Il modo in cui viene fatto questo è inserire una penalità *group-LASSO* che agisca per variabile, cioè ogni gruppo avrà al suo interno tutti i parametri che fanno riferimento alla p-esima variabile per ogni tempo t per $t = 1 \dots T$. Inoltre viene inserito un parametro adattivo per gruppo definito come $G_j = \frac{G}{|\boldsymbol{\beta}_j^{MQ}|}$ in cui β_j^{MQ} rappresenta il coefficiente relativo alla variabile j-esima con j=1,...,p calcolato come se le variabili non fossero dipendenti dal tempo, in pratica una regressione lineare statica sulla matrice $\mathbf{X}$. Il modello sarà così specificato:

$$\min_{\boldsymbol{\beta}} : \frac{1}{2}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \frac{\alpha}{2}\lambda\boldsymbol{\beta}^{\intercal}\boldsymbol{F_F}^{\intercal}\boldsymbol{F_F}\boldsymbol{\beta} + (1-\alpha)\lambda||\boldsymbol{F_F}\boldsymbol{\beta}||_1 + G\sum_{j=1}^{p}||\frac{\boldsymbol{F_j}\boldsymbol{\beta}}{\beta_p^{MQ}}||_2, \tag{3.63}$$

con $\boldsymbol{F_F}$ uguale a (3.8) e $\boldsymbol{F_j}$ che seleziona, come detto in precedenza, tutti i tempi della variabile j-esima. Il modo di stimare questo modello apparentemente molto complesso è in realtà un estensione dell'algoritmo *ADMM* visto per il caso (3.58). Sarà sufficiente aggiungere al processo di stima per il modello in equazione (3.58) un ulteriore passo

per $\boldsymbol{z}$ riguardante la stima dell'effetto della penalità *fused*. Si definisce inanzitutto il problema vincolato:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ \frac{1}{2}||\boldsymbol{y}-\mathbf{Z}\boldsymbol{\beta}||_2^2+\frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\mathbf{F}_F^\intercal\mathbf{F}_F\boldsymbol{\beta}+(1-\alpha)\lambda||\boldsymbol{z}_F||_1+G\sum_{j=1}^{p}||\frac{\boldsymbol{z}_j}{\beta_j^{MQ}}||_2 \\ &\textit{sotto il vincolo}: \mathbf{F}_i\boldsymbol{\beta}-\boldsymbol{z}_i=0 \qquad \textit{per} \quad i\in\{1,\dots,p,F\}, \end{aligned} \tag{3.64}$$

con $G\in\mathbb{R}^+$. Quindi la prima parte sarà identica al modello visto precedentemente mentre la seconda parte, che si configura come una penalità *group-LASSO* con in aggiunta i pesi $\frac{1}{|\beta_p^{MQ}|}$ che però non incidono nella stima in quanto in questa fase sono considerati noti. Quindi unendo i vari processi di stima visti nelle scorse sezioni si ottiene:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= (\mathbf{Z}^\intercal\mathbf{Z}+\gamma\mathbf{F}^\intercal\mathbf{F})^{-1}(\mathbf{Z}^\intercal\boldsymbol{y}+\rho\mathbf{F}^\intercal(\boldsymbol{z}^k-\tilde{\boldsymbol{u}}^k)) \\ \boldsymbol{z}_F^{k+1} &= S_{\frac{(1-\alpha)\lambda}{\rho}}(\mathbf{F}_F\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}_F^k) \\ \boldsymbol{z}_j^{k+1} &= S^*_{\frac{G_j}{\rho}}(\mathbf{F}_j\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}_j^k) \qquad\qquad \textit{per } j=1,\dots,p \\ \tilde{\boldsymbol{u}}^{k+1} &= \tilde{\boldsymbol{u}}^k+\mathbf{F}\boldsymbol{\beta}^{k+1}-\boldsymbol{z}^{k+1}, \end{aligned} \tag{3.65}$$

in cui $G_j=\frac{G}{|\beta_j^{MQ}|}$. Come detto in precedenza questo modello permette di sparsificare la stima in un contesto dinamico, purtroppo però è stato necessario aggiungere un altro parametro che porta quasi all'impossibilità di definire un trio di parametri (λ,α,G) ottimi o vicini ad esserlo. L'aggiunta dell'adattività e della terza penalità rende difficile la stima dei gradi di libertà. Di seguito si confronta il modello in equazione (3.63) con quello in equazione (3.46) lasciando fissi i parametri α e λ e variando il parametro G da 0.0005 a 2 in scala logaritmica, su dati simulati con struttura simile a quella vista nella sezione precedente. In particolare si sono simulati 200 osservazioni $y_i=\boldsymbol{x}_i^\intercal\boldsymbol{\beta}+\epsilon_i$ con $\boldsymbol{x}_i\sim N_p(0_p,\mathbf{I}_p)$, $\epsilon_i\sim N(0,1)$, $\beta_j\sim U(-5,5)$ con perturbazioni provenienti da $U(-10,10)$, ed infine con 4 su 10 coefficienti pari a 0 per tutta la dinamica, in modo da esplorare l'effetto della nuova penalità.

Come è possibile notare dalla figura (3.9) l'aumentare di G porta la stime verso lo zero traslando la dinamica dei coefficienti ma permette anche, quando i veri coefficienti sono pari a 0, di stimare tutta la dinamica di un determinato coefficiente a 0 di fatto sparsificando la stima dei coefficienti dinamici. Si mostra infine un dettaglio sui grafici in cui i coefficienti simulati sono pari a 0 per vedere nel dettaglio l'effetto di G.

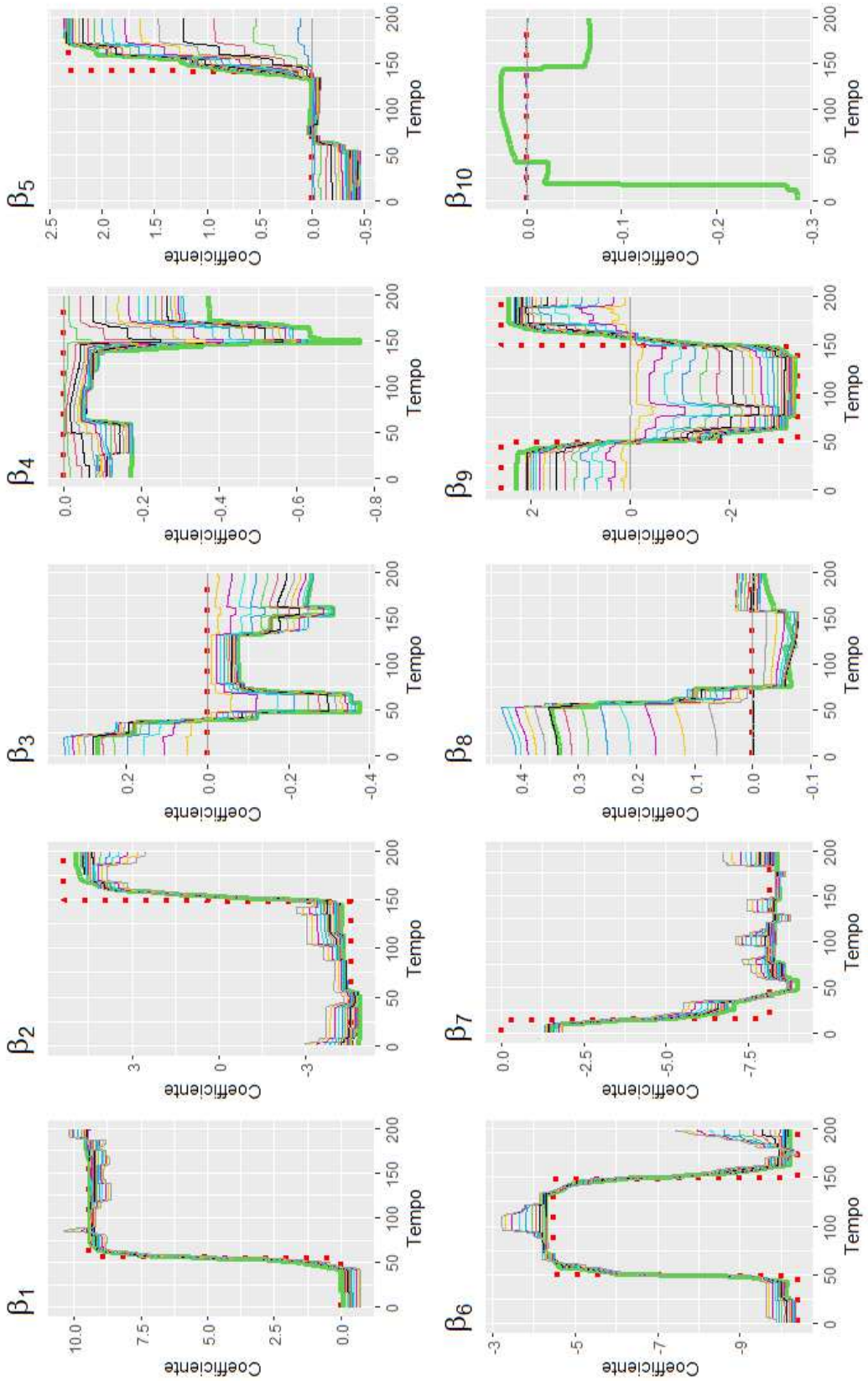

FIGURA 3.9: Coefficienti simulati (linea tratteggiata rossa), coefficienti stimati dal modello in equazione (3.46) (linea verde), coefficienti modello in equazione (3.63) al variare di G (linee colorate sottili).

È possibile notare dalla figura (3.10) come molti dei coefficienti siano stimati molto vicini allo 0 per vari valori di G, potendo così concludere che la penalità aggiunta dal modello (3.63), anche se a discapito della precisione degli altri coefficienti, raggiunge l'obbiettivo desiderato sparsificando la stima.

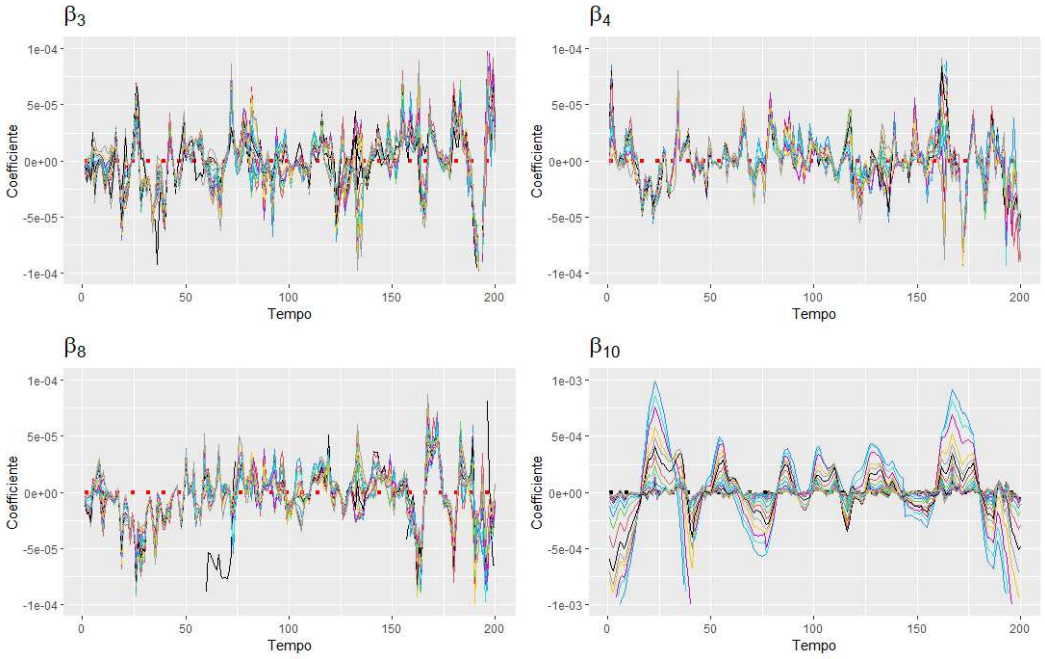

FIGURA 3.10: Coefficienti simulati (linea tratteggiata rossa), coefficienti modello in equazione (3.63) al variare di G (linee colorate sottili).

Infine è di interesse catturare l'effetto dell'aggiunta della penalità adattiva per effetto di ω_j. Si mostra quindi di seguito l'entrata dei vari coefficienti al diminuire di G con e senza i pesi ω_j. I dati utilizzati per questo esempio sono stati simulati similmente a quanto visto nel caso precedente, con un numero di osservazioni pari a 200 e 5 coefficienti tra cui 2 di questi uguali a 0 per tutta la dinamica. Come è possibile notare dai due grafici in figura (3.11) la sparsificazione dei due gruppi di parametri i quali $\boldsymbol{\beta}$ simulati sono uguali a 0 ($\boldsymbol{\beta}_2$ *e* $\boldsymbol{\beta}_4$) escono dal modello per G più piccoli nel modello adattivo piuttosto che in quello non adattivo. Inoltre per quanto riguarda i restanti tre coefficienti diversi da zero notiamo invece non solo un entrata nel modello è molto più distanziata nel modello adattivo piuttosto che nel modello non adattivo ma anche il fatto che essi escano dal modello per G molto più grandi. L'inserimento quindi della penalità adattiva potrebbe permette una più accurata selezione del modello.

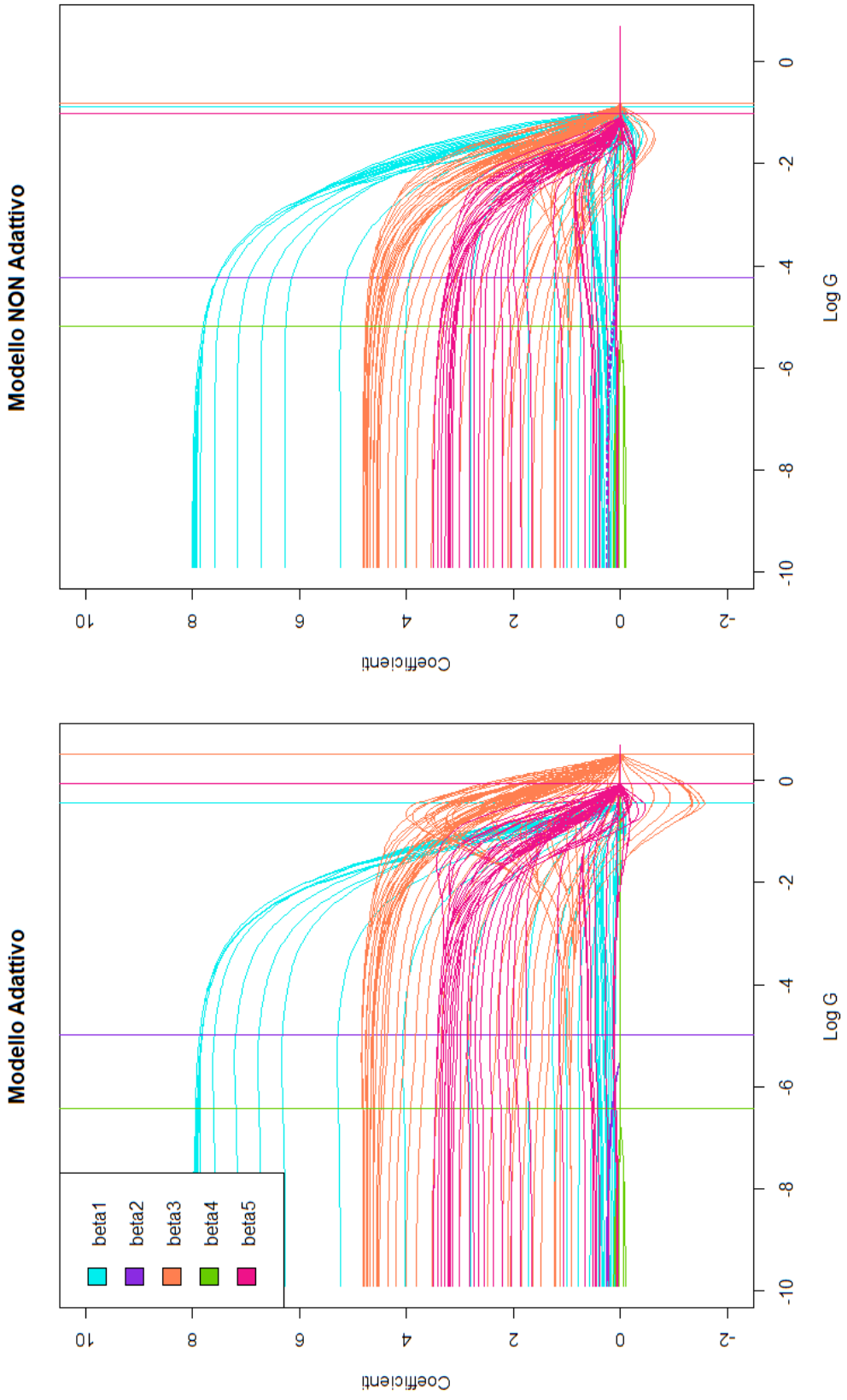

FIGURA 3.11: Percorso coefficienti al variare di *log(G)*.

Infine per i dati simulati per l'esempio di figura (3.9) si mostra il modello con il MAE minore tra quelli stimati a confronto con il modello in equazione (3.63):

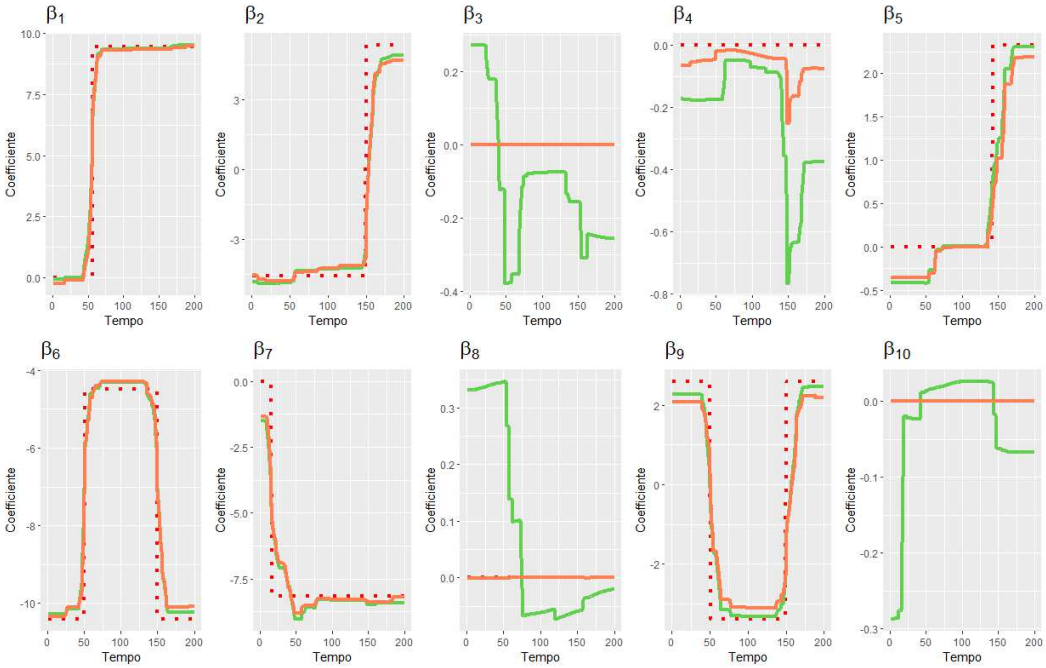

FIGURA 3.12: Coefficiente simulati (linea tratteggiata rossa), coefficienti modello in equazione (3.46) (linea verde), coefficienti ottimi modello in equazione (3.63) (linea corallo).

Si nota come nonostante una piccola traslazione e conseguente perdita di precisione riguardante i coefficienti diversi da zero, la sparsificazione della stima migliora molto la precisione per i coefficienti simulati uguali a 0. È poi di interesse focalizzarsi sul coefficiente $\boldsymbol{\beta}_4$ il quale non è stimato uguale a zero nella soluzione con il minore MAE, questo sottolinea il *trade-off* accennato prima, quindi per poter stimare anche quel coefficiente uguale a zero la perdita di precisione negli altri è maggiore del benefit portato dalla sua sparsificazione.

3.12 Fused elastic net con group LASSO per il tempo

Infine è di interesse introdurre un ultima penalità che agisca nei casi in cui il sistema per un periodo limitato di tempo si "spenga", quindi si introduce una penalità *group LASSO* che agisca per tempo, cioè i gruppi saranno rappresentati da variabili diverse appartenenti allo stesso istante di tempo, in questo modo queste potranno uscire dal modello contemporaneamente al tempo t. Questo è un caso comunque particolare che solitamente è difficile osservare. Il modello sarà così definito :

$$\begin{aligned}\min_{\boldsymbol{\beta}} :&\frac{1}{2}||\boldsymbol{y}-\mathbf{Z}\boldsymbol{\beta}||_2^2+\frac{\alpha}{2}\lambda\boldsymbol{\beta}^\intercal\boldsymbol{F_F}^\intercal\boldsymbol{F_F}\boldsymbol{\beta}+(1-\alpha)\lambda||\boldsymbol{F_F}\boldsymbol{\beta}||_1+ \\ &G_1\sum_{j=1}^{p}||\frac{\boldsymbol{F_j}\boldsymbol{\beta}}{\beta_j^{MQ}}||_2+G_2\sum_{t=1}^{\intercal}||\boldsymbol{F_t}\boldsymbol{\beta}||_2,\end{aligned} \tag{3.66}$$

con $\mathbf{F}_t$ matrice che seleziona le p variabili al tempo t. Come in precedenza si decide di aggiungere un ulteriore parametro di penalità (G_2) che non sia in competizione con quello precedentemente introdotto (qui chiamato G_1) in quanto riguardano due fenomeni diversi e difficilemente saranno presenti entrambe le penalità nel modello. Anche qui è difficile trovare un metodo per regolare i 4 parametri di penalità. Il metodo di stima comunque non si complica particolarmente infatti basterà specificare anche in questo caso un ulteriore passo per $\mathbf{Z}$ che stimi l'effetto dell'ultima penalità. L'*ADMM* diventerà quindi:

$$\begin{aligned}\boldsymbol{\beta}^{k+1}&=(\mathbf{Z}^\intercal\mathbf{Z}+\gamma\mathbf{F}^\intercal\mathbf{F})^{-1}(\mathbf{Z}^\intercal\boldsymbol{y}+\rho\mathbf{F}^\intercal(\boldsymbol{z}^k-\tilde{\boldsymbol{u}}^k)) \\ \boldsymbol{z}_F^{k+1}&=S_{\frac{(1-\alpha)\lambda}{\rho}}(\mathbf{F}_F\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}_F^k) \\ \boldsymbol{z}_j^{k+1}&=S^*_{\frac{G_{1,j}}{\rho}}(\mathbf{F}_j\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}_j^k) \qquad per\ j=1,\dots,p \\ \boldsymbol{z}_t^{k+1}&=S^*_{\frac{G_2}{\rho}}(\mathbf{F}_t\boldsymbol{\beta}^{k+1}+\tilde{\boldsymbol{u}}_t^k) \qquad per\ t=1,\dots,T \\ \tilde{\boldsymbol{u}}^{k+1}&=\tilde{\boldsymbol{u}}^k+\mathbf{F}\boldsymbol{\beta}^{k+1}-\boldsymbol{z}^{k+1}.\end{aligned} \tag{3.67}$$

Si confronta di seguito il percorso dei coefficienti al variare di G_2 per il modello in equazione (3.66) con G_1=0 con quello in equazione (3.46) con (λ,α) fissati (uguali per entrambi i modelli), su dati simulati con T=200, p=5 e i $\boldsymbol{\beta}$ simulati che per un intervallo di tempo casuale diventano uguali a 0 contemporaneamente.

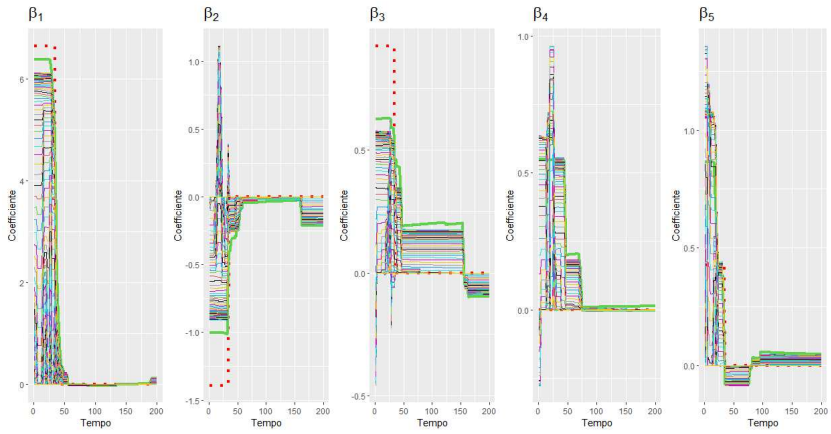

FIGURA 3.13: Coefficienti simulati (linea rossa trateggiata), coefficienti modello (3.60) (linea verde grossa), coefficienti modello (3.63) (linee sottili).

Si può notare dalla figura (3.13) come l'introduzione della penalità permetta di sparsificare la dinamica dei coefficienti per determinati tempi ma danneggia la sparsificazione comprensiva della stima, in quanto per gli intervalli di tempi per cui alcuni dei coefficienti sono diversi da zero ed altri no la stima tenderà o a stringere troppo i coefficienti o, al contrario, a stimarli diversi da zero. Questo si può vedere ad esempio nel grafico di $\boldsymbol{\beta}_5$ le stime spingono a stimare coefficienti diversi da 0 a causa del fatto che gli altri coefficienti sono presenti nel modello simulato. Si confronta in figura (3.14), il modello con il *MAE* minore tra quelli stimati e il modello in equazione (3.46) evidenzionando l'effetto della penalità a scapito sparsificazione omogenea di alcuni coefficienti.

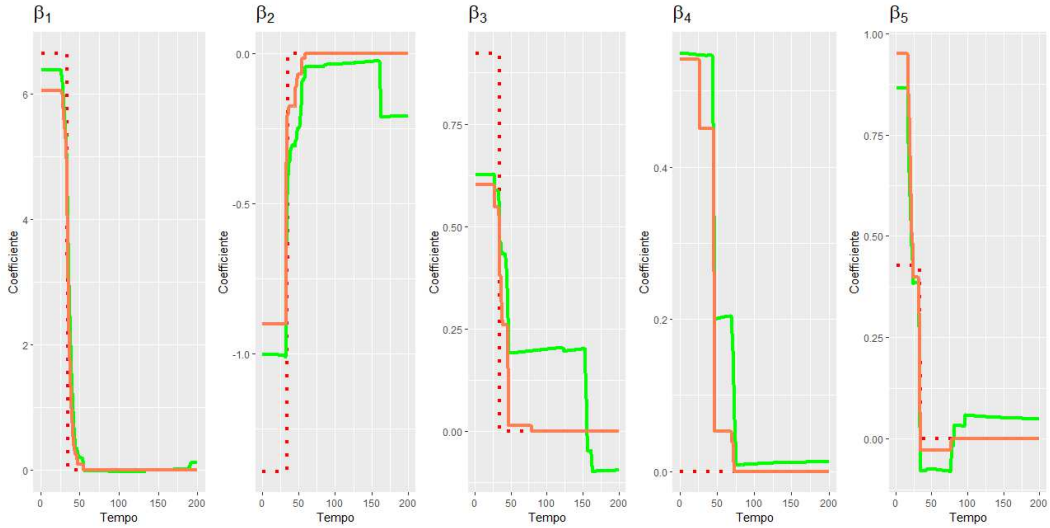

FIGURA 3.14: Coefficienti simulati (linea rossa trateggiata), coefficienti modello in equazione (3.46) (linea verde), coefficienti modello ottimo in equazione (3.66) (linea corallo).

Infine lasciando sempre fisse (λ, α) si utilizza una griglia di (G_1, G_2) per esplorare l'effetto congiunto delle due penalità introdotte. Si utilizzano dati simulati per 200 osservazioni e 10 variabili con modalità uguali a quelle utilizzate per l'esempio precedente. Si utilizza una griglia di 10 valori in scala logaritmica per entrambi i parametri di regolazione portando a 100 modelli stimati. Si mostrano in figura (3.15) il percorso dei vari coefficienti.

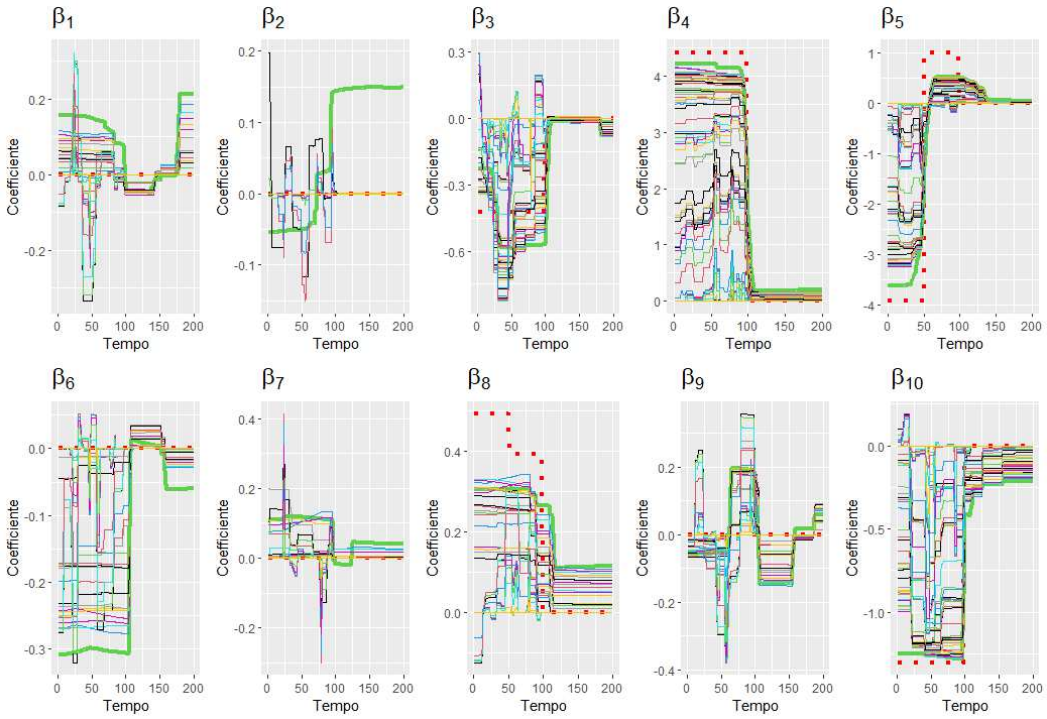

FIGURA 3.15: Coefficienti simulati (linea rossa trateggiata), coefficienti modello (3.60) (linea verde), coefficienti modello al variare di G_1 e G_2 (3.69) (linee colorate sottili).

la prima cosa da notare nella figura (3.15) è la maggiore erraticità delle stime appunto dovuta al fatto che si hanno due parametri che variano invece di uno solo. Per maggiore chiarezza, in figura (3.16), si analizzano alcune stime specifiche per alcuni valori specifici di G_1 e G_2. Si nota come nel primo caso (valore estremo per G_1), tutti i coefficienti vengono appiattiti a 0, nel secondo caso invece (valore estremo per G_2) si riesce a stimare la seconda parte dei coefficienti uguali a 0 ma i restanti portano delle stime erratiche ed infine nel caso intermedio si riesce a stimare tutti i coefficienti simulati uguali a 0 efffettivamente tali e in parte quelli che hanno solo una parte uguale a 0 riuscendo, anche se con poca precisione, quelli diversi da 0. Come si era ipotizzato quindi la presenza di entrambe le penalità contemporaneamente rende il modello molto instabile e non molto utile senza una conoscenza a priori del fenomeno.

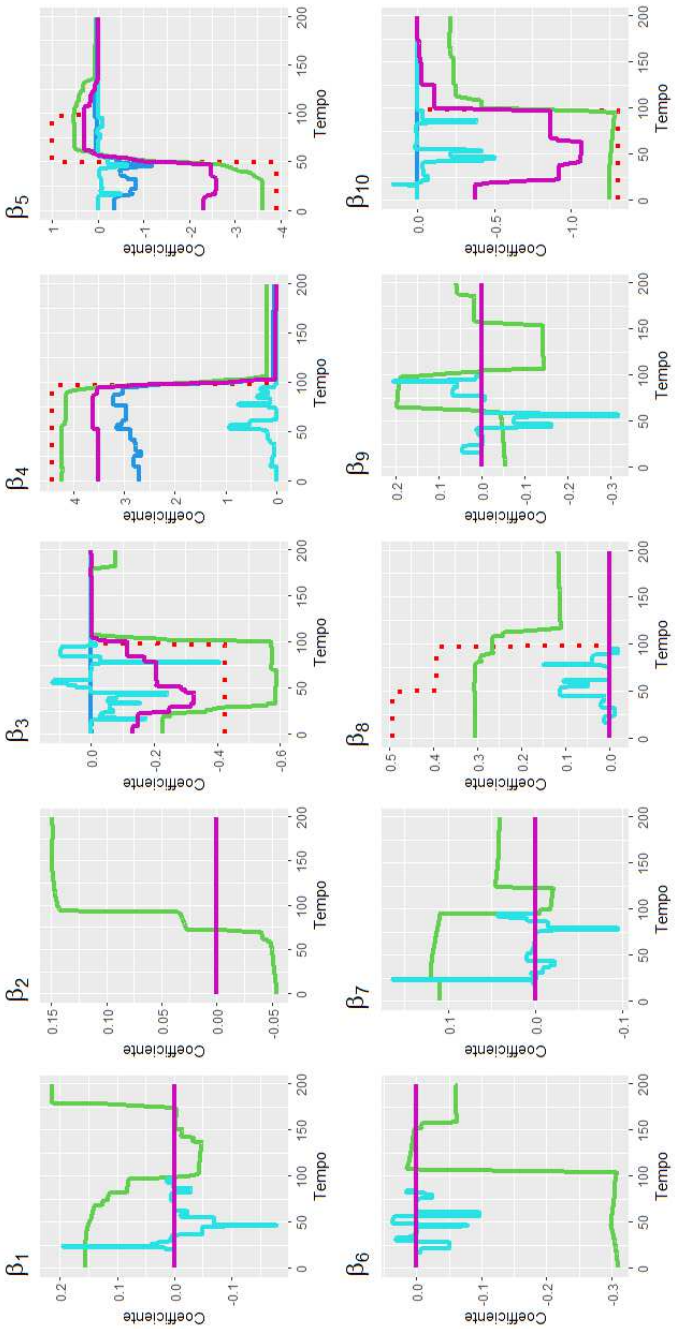

FIGURA 3.16: Coefficienti simulati (linea rossa trateggiata), coefficienti modello (3.60) (linea verde), coefficienti modello (3.69) G_1=0.075 e G_2=0.0066 (linea blu), coefficienti modello (3.69) G_1=0.0066 e G_2=0.075 (linea azzurra), coefficienti modello (3.69) G_1=0.011 e G_2=0.0017 (linea viola).

Capitolo 4

Applicazione: struttura della curva dei rendimenti

Nel seguito si tratterà il tema della curva dei rendimenti. In finanza, titoli di credito (titoli di stato ad esempio) con caratteristiche identiche ma con diverse scadenze hanno tassi di rendimento diversi a causa dell'incertezza data dal tempo in termini di rischio: in situazioni normali più lunga è la scadenza di un titolo più alto è il suo rendimento. È quindi possibile tracciare una curva definendo i tassi di rendimento per un certo titolo ad un certo istante di tempo per diverse scadenze. Oltre alla stima della relazione esistente tra i rendimenti ed il tempo, in questo campo è d'interesse anche la previsione dei tassi futuri basandosi sulle informazioni del passato.

4.1 Stima e previsione della curva dei rendimenti

In letteratura il modello più famoso per trattare questo tema è quello proposto da Nelson and Siegel (1987) che definiscono la curva dei rendimenti come :

$$y_t(\tau) = \boldsymbol{\beta}_{1t} + \boldsymbol{\beta}_{2t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau}\right) + \boldsymbol{\beta}_{3t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau} - e^{-\lambda_t \tau}\right). \tag{4.1}$$

Come descritto in Diebold and Li (2006), λ_t guida il tasso di decadimento esponenziale dei tassi, quindi valori alti di λ_t portano ad un decadimento lento e meglio si adattano a scadenze più lunghe, mentre valori grandi di λ_t producono un decadimento veloce che meglio si adatta a scadenze più brevi. Infine, questo parametro definisce anche il massimo del fattore che fa riferimento al parametro $\boldsymbol{\beta}_{3t}$. Per quanto riguarda $\boldsymbol{\beta}_{1t}$ questo rappresenta il parametro di lungo periodo in quanto il suo fattore di riferimento è uguale a 1 e di conseguenza non decade all'aumentare di τ. I fattori che fanno riferimento a $\boldsymbol{\beta}_{2t}$ e

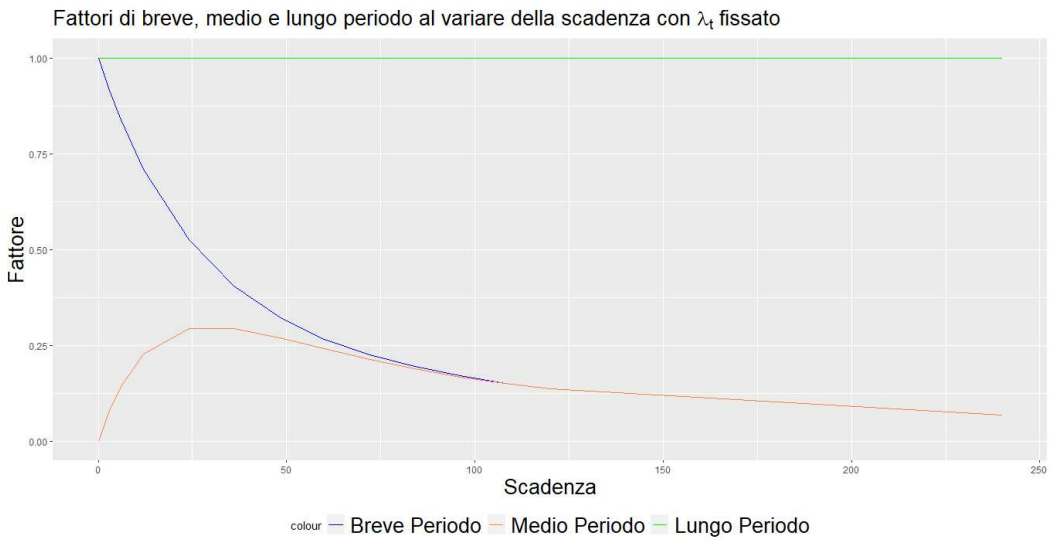

FIGURA 4.1: Fattori al variare di τ.

$\boldsymbol{\beta}_{3t}$ rappresentano rispettivamente il fattore di breve e di medio periodo; infatti $(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau})$ è una funzione che parte da 1 e decade velocemente e in maniera monotona verso lo 0, mentre $(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau})$ parte da 0, aumentando inizialmente per poi raggiungere un massimo e successivamente decadere verso lo zero. Inoltre, questi in lettaratura sono anche interpretati come il livello, l'inclinazione e la curvatura della curva. In Diebold and Li (2006) si specifica che un buon modello riguardante la dinamica della curva dei rendimenti dovrebbe essere capace di riprodurre la dinamica storica riguardante la media della curva dei rendimenti, evidenziare le varie forme che essa assume nei diversi tempi ed infine descrivere la persistenza dei rendimenti e la non persistenza degli spreads. Nel framework proposto in Diebold and Li (2006) per stimare i parametri del modello (4.1) viene fissato λ_t massimizzando il fattore di medio-termine a 30 mesi ottenendo $\widehat{\lambda_t}$ =0.0609 e di seguito viene minimizzata la somma dei quadrati della distanza tra i rendimenti $y_t(\tau)$ osservati e quelli stimati dall'equazione (4.1). Questo viene fatto tempo per tempo quindi di base non è presente una dinamica tra i parametri ma si otterrano 3 serie storiche una per parametro$(\boldsymbol{\beta}_{1t}, \boldsymbol{\beta}_{2t}, \boldsymbol{\beta}_{3t})$.

In questo elaborato quindi si propone di applicare il modello (3.46) al problema della curva dei rendimenti andando a definire un modello di regressione come segue:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \epsilon_t, \tag{4.2}$$

con $\mathbf{Z} = \mathbf{I}_T \otimes \mathbf{Z}_t$ matrice diagonale a blocchi di dimensioni $13T \times 3T$, con T numero di osservazioni e con elementi diagonali:

$$\mathbf{Z}_t = \begin{bmatrix} 1 & (\frac{1-e^{-\widehat{\lambda}_t 3}}{\widehat{\lambda}_t 3}) & (\frac{1-e^{-\widehat{\lambda}_t 3}}{\widehat{\lambda}_t 3} - e^{-\widehat{\lambda}_t 3}) \\ 1 & (\frac{1-e^{-\widehat{\lambda}_t 6}}{\widehat{\lambda}_t 6}) & (\frac{1-e^{-\widehat{\lambda}_t 6}}{\widehat{\lambda}_t 6} - e^{-\widehat{\lambda}_t 6}) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & (\frac{1-e^{-\widehat{\lambda}_t 240}}{\widehat{\lambda}_t 240}) & (\frac{1-e^{-\widehat{\lambda}_t 240}}{\widehat{\lambda}_t 240} - e^{-\widehat{\lambda}_t 240}) \end{bmatrix} \quad \in \quad \mathbb{R}^{13\times 3}, \tag{4.3}$$

con $\boldsymbol{y}$ che impila i rendimenti $(y_t(3), y_t(6), \ldots, y_t(240))$.

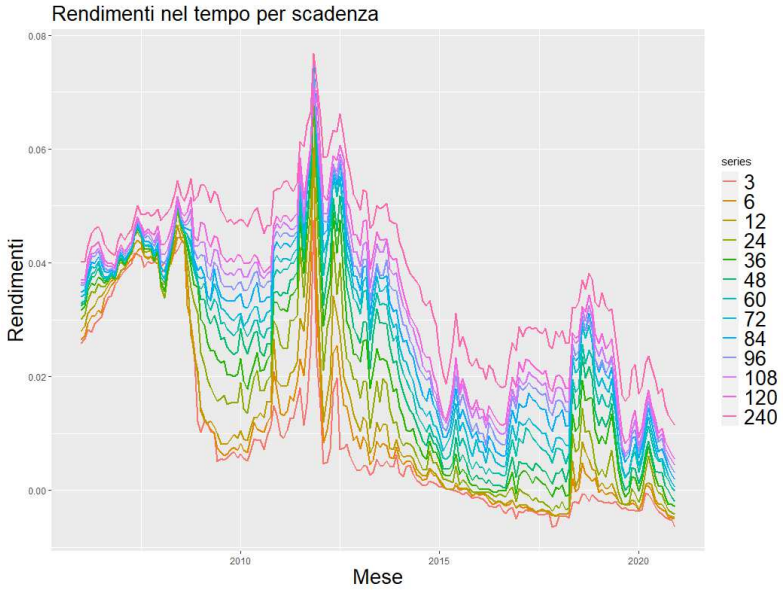

FIGURA 4.2: Rendimenti nel tempo per varie scadenze dal 2006 al 2022.

Si utilizzeranno i rendimenti dei buoni del tesoro dello stato italiano dal 2006 al 2022 senza cedola con osservazioni mensili per un totale di 204 tempi, per ogni tempo si avranno quindi 13 rendimenti che descrivono il rendimento atteso per varie scadenze. Per questa applicazione, per la selezione del modello in equazione (3.46) si utilizzerà una sequenza di λ di 100 valori da 0.0001 all'autovalore massimo di $\mathbf{Z}^\intercal\mathbf{Z}$ (15.43) e una sequenza di α di 5 valori da 0.1 a 0.9. Inoltre, in questo caso si utilizzerà il criterio di informazione BIC per accentuare l'effetto della penalità. Il parametro λ_t viene fissato a 0.0609 per tutti i tempi, questo valore è trovato come accennato precedentemente massimizzando il fattore di medio periodo a 30 mesi:

$$\widehat{\lambda} = \max_{\lambda}\left(\frac{1-e^{-\lambda 30}}{\lambda 30} - e^{-\lambda 30}\right). \tag{4.4}$$

Come si può notare dalla figura 4.2 all'aumentare della scadenza i tassi di rendimento aumentano in quanto la variabile tempo crea incertezza e questo si riflette come un premio nel rendimento. In figura (4.3) si mostrano i valori del BIC al variare di α e λ (figura 3.19) calcolati utilizzando l'approssimazione per la complessità del modello in equazione (3.57).

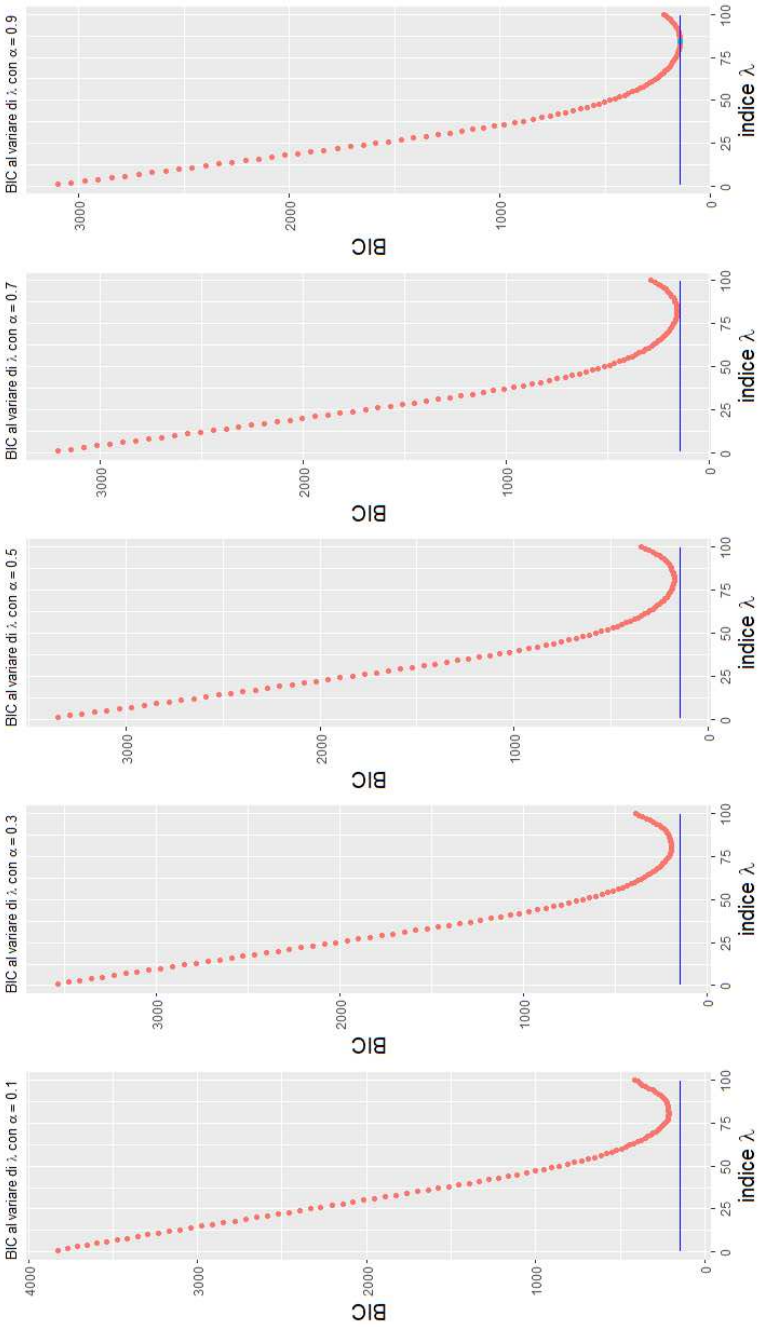

FIGURA 4.3: BIC al variare di α e λ.

Si nota come tutti gli andamenti sono coerenti con quanto ci si aspetta da un criterio di informazione, inoltre in questo caso, differentemente dallo studio di simulazione, la matrice di disegno è 2695×612 ed è quindi a rango pieno. Questo risultato non è banale in quanto mostra l'efficacia dell'approssimazione di d eseguita tramite la funzione maggiorante in equazione (3.54). Il valore ottimo per la coppia (λ, α) è stimata come $(2.24,\ 0.9)$. Si mostrano di seguito due curve dei rendimenti stimate dal modello in equazione (3.46) con due comportamenti diversi.

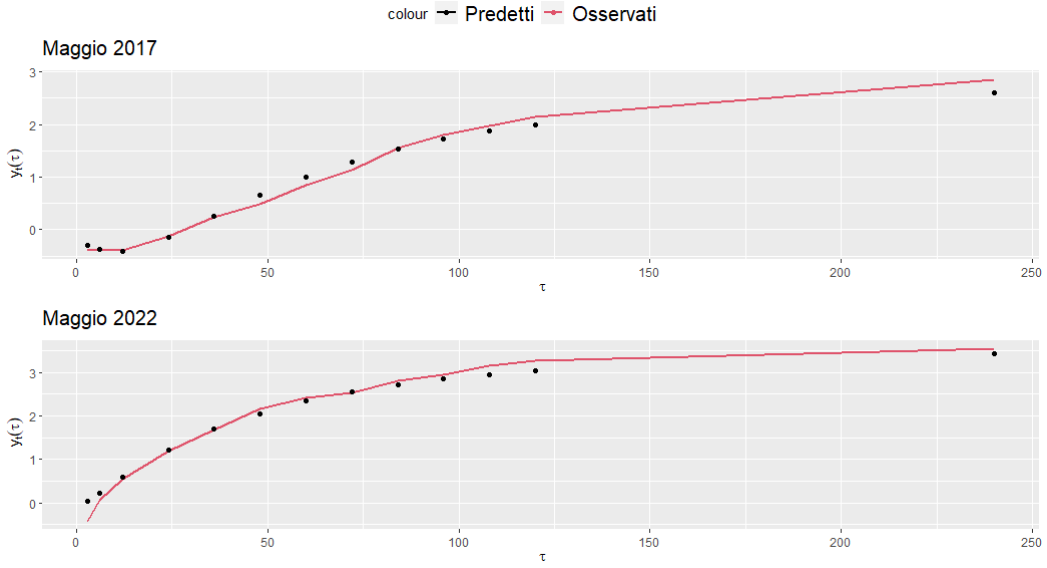

FIGURA 4.4: Curve dei rendimenti per due date specifiche.

Per quanto riguarda il modello Nelson Siegel in Diebold and Li (2006) questo viene stimato attraverso la minimizzazione della somma al quadrato delle distanze $y_t(\tau) - \widehat{y}_t(\tau)$ in cui $\widehat{y}_t(\tau) = \widehat{\boldsymbol{\beta}}_{1t} + \widehat{\boldsymbol{\beta}}_{2t}(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}) + \widehat{\boldsymbol{\beta}}_{3t}(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau})$, questo però viene fatto tempo per tempo, cioè i singoli $\widehat{\boldsymbol{\beta}}_{it}$ saranno indipendenti tra loro. Successivamente quindi per creare una dinamica di dipendenza temporale viene stimato un modello AR(1) per i tre $\boldsymbol{\beta}_{it}$. In figura 4.5 vengono quindi mostrati i risultati della stima classica e del modello in equazione (3.46) il quale, invece, data la struttura della stima, ha già al suo interno la stima di coefficienti dinamici. Come si può notare, il modello in equazione (3.46) media molto bene la stima dei coefficienti rendendo le tre serie meno variabili ed inoltre per $\boldsymbol{\beta}_3$ il coefficiente resta sempre negativo per tutto la dinamica, cosa che invece non accade nella stima classica a causa di una grossa varibilità presente intorno al 2013.

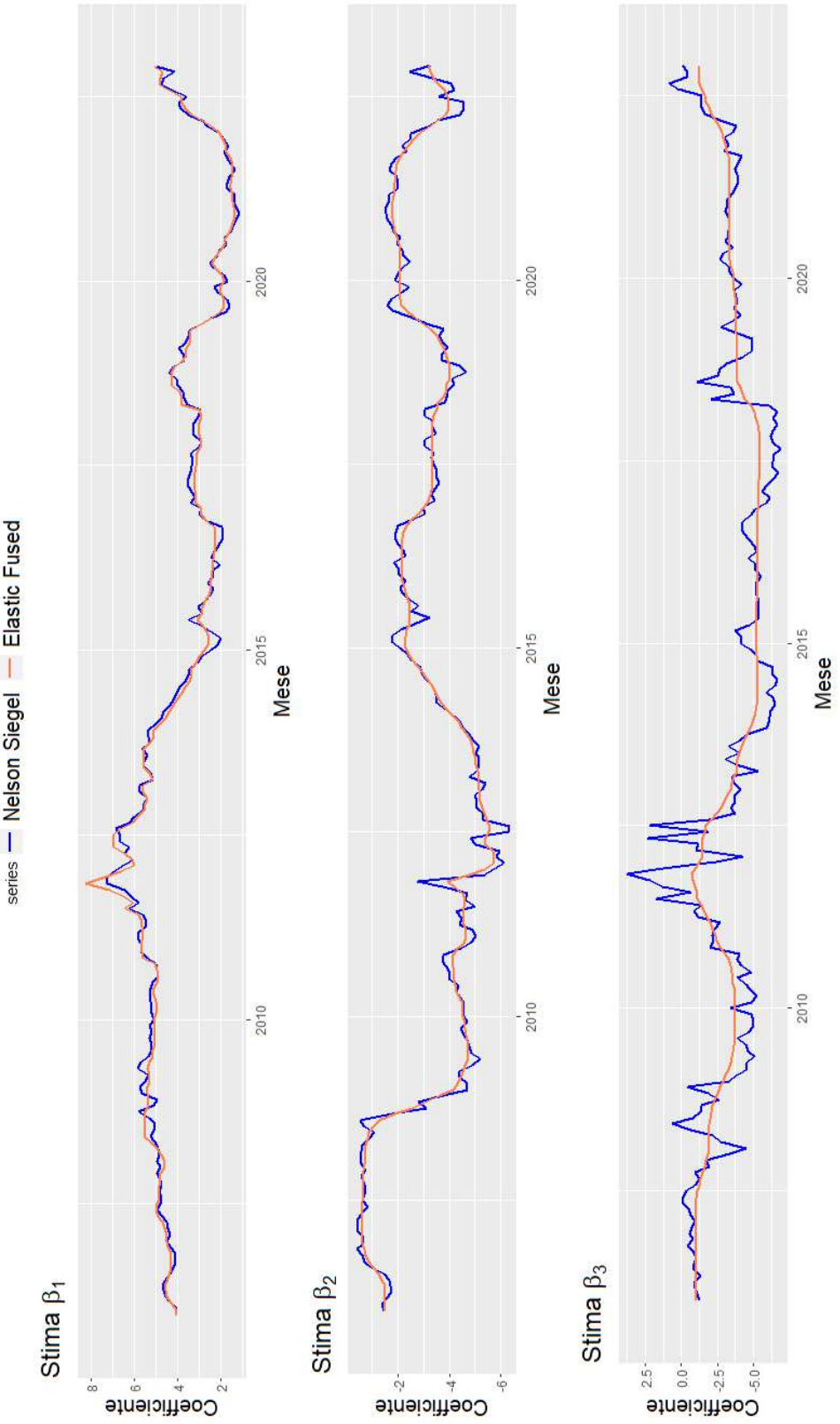

FIGURA 4.5: Stime coefficienti modello in equazione (4.1).

Si mostrano di seguito i valori predetti dai due modelli per i vari rendimenti assieme ai valori osservati.

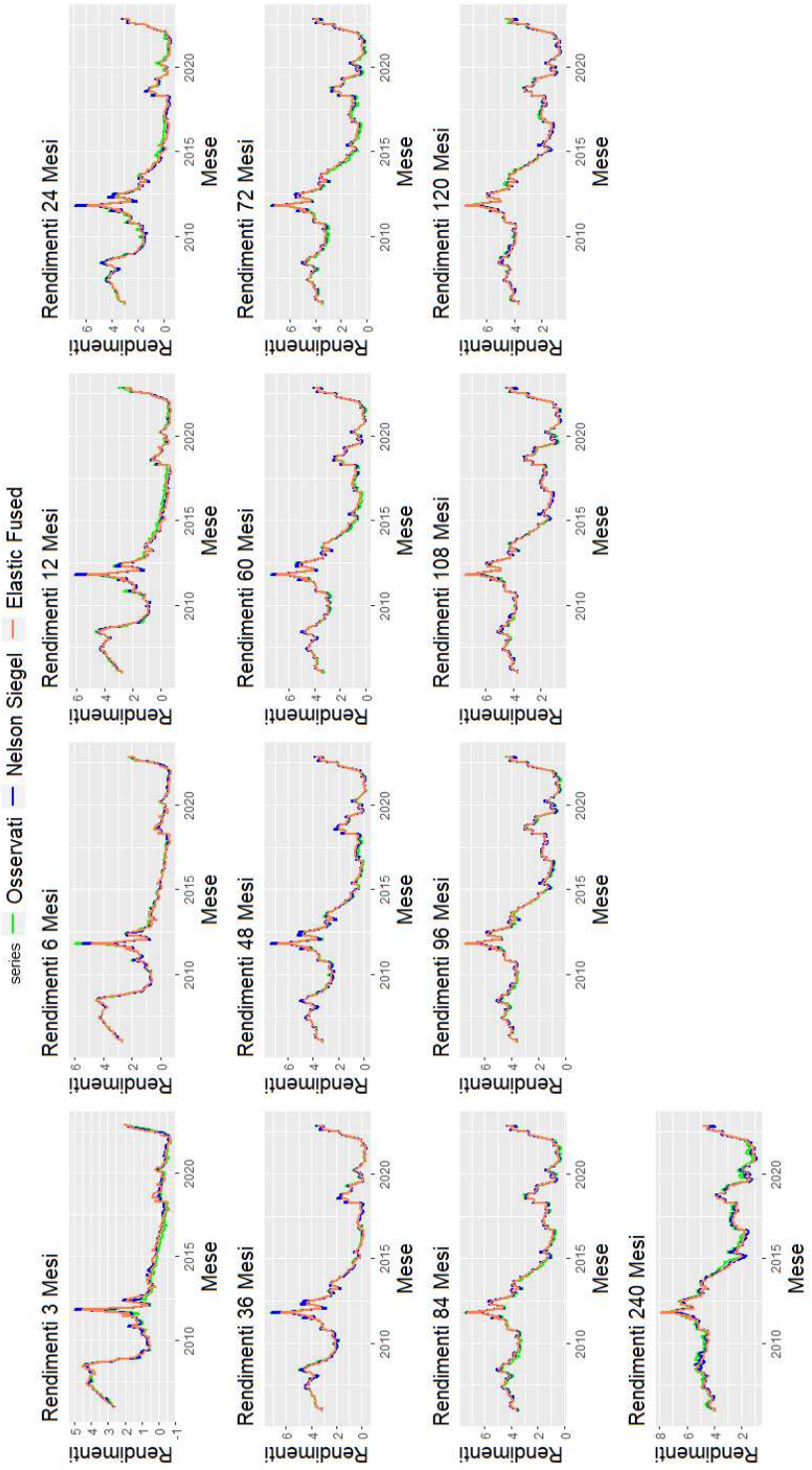

FIGURA 4.6: Rendimenti previsti e osservati.

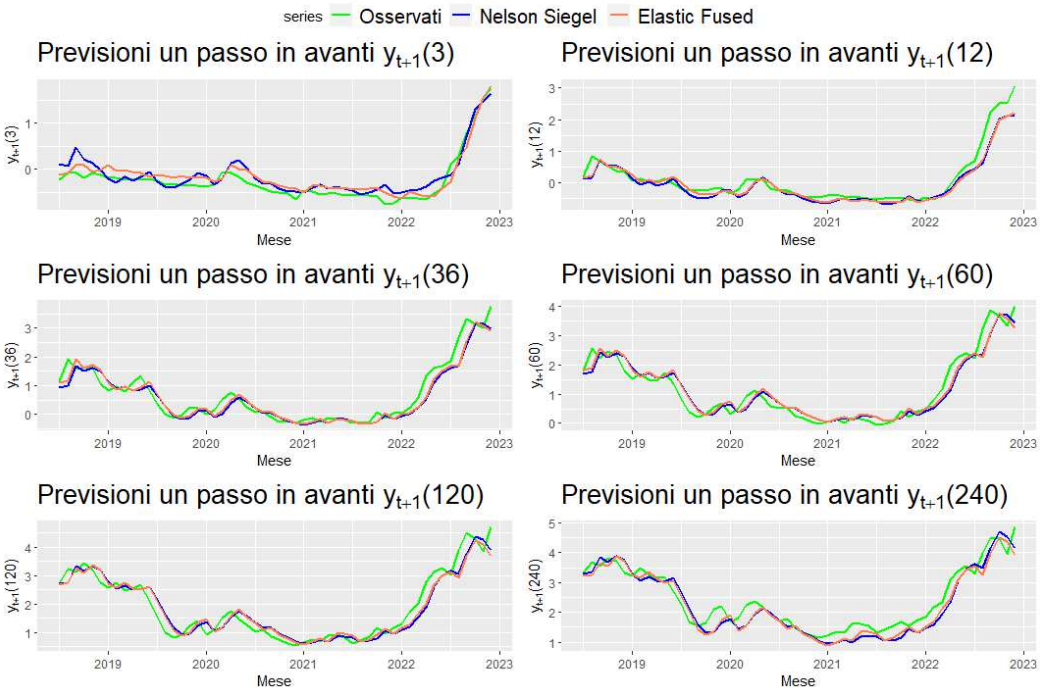

FIGURA 4.7: Confronto previsioni un passo in avanti per i due modelli.

Infine, si valuta la precisione della previsione dei tassi di rendimento in una finestra *rolling*:

- nel caso Nelson Siegel si stimerà ricorsivamente partendo dalla 150-esima osservazione un modello AR(1) per $\boldsymbol{\beta_{it}}$ (per i= 1,2,3) e si predirrà $\boldsymbol{\beta_{iT+1}}$ e di conseguenza $\widehat{y}_{T+1}(\tau)$ per τ=(3,12,36,60,120,240)

- per il modello *elastic-fused dinamico* (3.48) si stimerà in maniera ricorsiva partenendo dalla 150-esima osservazione e per ogni tempo si utilizzerà $\boldsymbol{\beta_{iT}}$ (con T=150 alla prima iterazione, 151 alla seconda e cosi via) come migliore stima di $\boldsymbol{\beta_{iT+1}}$ e di conseguenza si stimerà $\widehat{y}_{t+1}(\tau)$ per τ=(3,12,36,60,120,240)

In totale quindi si avranno 54 previsioni e 54 modelli che saranno utilizzati per la stima dei rendimenti.

Come è possibile notare dalla figura (4.7) le prevsioni dei due modelli sono molto simili e questo giustifica quindi l'uso del modello proposto (3.46) anche in campo previsionale.

Si confrontano quindi di seguito le previsioni un passo avanti dei due modelli in termini di media degli errori $(y_{t+1}(\tau) - \widehat{y}_{t+1}(\tau))$, RMSE (radice quadrata dell'errore medio al quadrato) e deviazione standard degli errori come proposto in Diebold and Li (2006).

TABELLA 4.1: *Errori di previsione un passo in avanti*

	Elastic-Fused			*Nelson Siegel AR(1)*		
Scadenza	Media	Deviazione standard	RMSE	Media	Deviazione standard	RMSE
3 Mesi	0.11	0.16	0.19	0.10	0.13	0.16
12 Mesi	-0.15	0.24	0.28	-0.12	0.24	0.27
36 Mesi	-0.13	0.32	0.35	-0.09	0.32	0.32
60 Mesi	-0.016	0.33	0.33	0.014	0.33	0.32
120 Mesi	-0.039	0.36	0.36	-0.035	0.36	0.36
240 Mesi	-0.17	0.33	0.37	-0.18	0.31	0.36

Come ci si poteva aspettare dopo la visione delle serie storiche riguardanti le previsioni ad un passo, le metriche di errore dei due modelli sono molto simili.

Conclusioni

In questo elaborato è stata descritta la flessibilità dell'algoritmo *ADMM* per la stima di modelli ad elevata dimensionalità penalizzati. In particolare il modello proposto, e le sue estensioni, possono essere stimate facilmente e con soluzioni abbastanza standard. Inoltre questa flessibilità dell'algoritmo permette, come visto, di utilizzare l'informazione a priori a disposizione per scegliere la penalità più adatta ed implementare così la soluzione per l'algoritmo *ADMM*. Il campo in cui ci si è spinti in questo elaborato resta comunque molto complesso, in primis per le dimensioni della matrice di disegno ma anche per la sua struttura rigida. Questo però ha permesso di ragionare sulla stima dei gradi di libertà del modello , argomento non sempre approfondito nei casi non standard in quanto di difficile trattamento. Sarebbe quindi interessante ampliare lo studio per il lisciamento di funzioni di perdita per la selezione del modello in ambiti in cui la struttura del modello non permette manipolazioni della matrice di disegno, un esempio possono essere alcuni modelli funzionali complessi. Un risultato da sottolineare riguarda l'applicazione vista nell'ultimo esempio in cui la matrice di disegno è a rango pieno: in questo caso il criterio di informazione utilizzato congiuntamente all'approsimazione dei gradi di libertà proposta in equazione (3.57) agiscono come ci si aspetta da una stima classica portando ad una buona selezione del modello. L'obiettivo di questa tesi è stato quindi quello di esplorare il mondo dei modelli penalizzati trovando forme e metodi di stima standard sia per modelli con strutture classiche che non, affrontandone i problemi riguardanti la selezione del modello e proponendo una via alternativa alla classica.

Per quanto riguarda invece la struttura dinamica del modello questa sicuramente è risultato un po‘ rigida rispetto alla struttura presentata dai modelli *state-space* anche se come mostrato questi hanno dei punti in comune. L'impossibilità di esplicitare in maniera diretta un modo per predirre le osservazioni future è sicuramente limitante, infatti come estensione di questo modello si potrebbe pensare di strutturare un problema risolvibile in maniera sequenzale sfruttando la flessibilità dell'algoritmo *ADMM* formulando il problema vincolato in maniera opportuna.

Appendice A

Codice C++

In questo appendice vengono riportati alcuni dei principali codici sviluppati in C++ con l'ausilio della libreria ARMADILLO 12.4.

A.1 Algoritmo ADMM per la stima del modello in equazione 3.68

```
# A              : matrice delle covariate
# F_f            : matrice delle differenze prime
# F_fTF_f        : quadrato della matrice delle differenze prime
# Z              : matrice degli fissi
# b              : vettore delle variabili risposta
# pesi           : vettore dei pesi per la penalita' lasso
# u              : vettore di inizializzazione per la variabile duale
# z              : vettore di inizializzazione per la variabile ausiliaria
# lambda_f       : parametro di regolazione penalit\'a fused
# alpha_f        : parametro di regolazione penalit\'a norma l_1 e norma l_2
# lambda_l_T     : parametro di regolazione penalit\'a group-LASSO per tempo
# lambda_l_p     : parametro di regolazione penalit\'a group-LASSO per variabile
# rho_relaxation : algoritmo con relaxation
# rho            : variabile per la convergenza dell'algoritmo
# tau            : parametro per l'adattivit\'a dell'algoritmo
# alpha          : parametro per l'adattivit\'a dell'algoritmo
# reltol         : soglia per l'errore relativo
# reltol         : soglia per l'errore assoluto
# maxiter        : numero massimo di iterazioni

Rcpp::List Elastic_Fused_Lasso2(const arma::mat& A,arma::mat& F_f,arma::mat& F_fTF_f,arma::mat& Z,
                          const arma::colvec& b, const vec pesi, arma::colvec& u, arma::colvec& z,
                          const double lambda_f,const double alpha_f,const double lambda_l_T,
                          const double lambda_l_p,bool rho_relaxation, double rho,
                          const double tau,
                          const double alpha,const double reltol, const double abstol,
                          const int maxiter) {
```

```
/* Defiizione varibile */
int k, g_id_init, g_id_end;
double elTime;

/* Dimensioni */
const int m      = A.n_rows;
const int n_cov  = Z.n_cols;
const int p      = A.n_cols;
const int n      = p/m;
double    sqrtm  = std::sqrt(static_cast<float>(m));

/* Definizione di vettori e matrici  */



/* Matrici */
arma::mat A_m(m,p);
arma::mat gamma(p,p,fill::zeros);
arma::mat Beta(m,n);
arma::mat eye_m(m, m, fill::eye);
arma::mat ZZ(n_cov, n_cov, fill::zeros);
arma::mat L_ZZ(n_cov, n_cov, fill::zeros);
arma::mat L_ZZ_INV(n_cov, n_cov, fill::zeros);
arma::mat ZZ_INV(n_cov, n_cov, fill::zeros);
arma::mat P_ZZ(m, m, fill::zeros);
arma::mat M_ZZ(m, m, fill::zeros);
arma::mat M_ZZW(m, p, fill::zeros);
arma::mat M_ZZW2(m, m, fill::zeros);
arma::mat WMW(p, p, fill::zeros);
arma::mat ZW(n_cov, p, fill::zeros);
arma::mat P_ZW(n_cov, p, fill::zeros);
arma::mat U(p, p, fill::zeros);
arma::mat L(p, p, fill::zeros);
arma::mat F((m-1)*n+p,p,fill::zeros);
arma::mat FTF(p, p, fill::zeros);
arma::mat F_l(p,p,fill::zeros);
arma::mat F_l_p(p,p,fill::eye);
arma::mat FFTFF(p,p,fill::zeros);

/* Vettori colonna */
arma::colvec v(n_cov, fill::zeros);
arma::colvec x(p, fill::zeros);
arma::colvec x_star(p, fill::zeros);
arma::colvec q(p, fill::zeros);
arma::colvec z_old(n*(2*m-1), fill::zeros);

/* Vettori */
arma::vec ones_m(m, fill::ones);
arma::vec WMy(n, fill::zeros);
arma::vec Zy(n_cov, fill::zeros);
arma::vec v_LS(n_cov, fill::zeros);
arma::vec Wy(p, fill::zeros);
arma::vec h_objval(maxiter, fill::zeros);
arma::vec h_r_norm(maxiter, fill::zeros);
```

```
arma::vec h_s_norm(maxiter, fill::zeros);
arma::vec h_eps_pri(maxiter, fill::zeros);
arma::vec h_eps_dual(maxiter, fill::zeros);
arma::vec rho_store(maxiter+1, fill::zeros);

/* Definizione Matrice che modellano le penalita'*/
F_l = get_lasso_g(m,n);
F=join_vert(F_f,F_l,F_l_p);

/* ::::::::::::::::::::::::::::::::::::::::::::
Tempo computazionale                              */
wall_clock timer;
timer.tic();

/* rho per caso relaxation */
rho_store(0) = rho;

/* Calcolo variabili preliminari */
Z         = Z.each_col() % (ones_m * (1.0 / sqrtm));
A_m       = A.each_col() % (ones_m * (1.0 / sqrtm));
ZZ        = Z.t() * Z;
L_ZZ      = (chol(ZZ)).t();
L_ZZ_INV  = inv(trimatl(L_ZZ));
ZZ_INV    = L_ZZ_INV.t() * L_ZZ_INV;
P_ZZ      = Z * ZZ_INV * Z.t();
M_ZZ      = eye_m - P_ZZ;
WMW       = A_m.t() * M_ZZ * A_m; //AT(Residui A-Zbeta_A)
WMy       = A_m.t() * M_ZZ * b / static_cast<float>(sqrtm); //AT(Residui y-Zbeta_y)
Zy        = Z.t() * b / static_cast<float>(sqrtm);
v_LS      = L_ZZ_INV.t() * L_ZZ_INV * Zy; // Beta_Zy
ZW        = Z.t() * A_m ;
P_ZW      = L_ZZ_INV.t() * L_ZZ_INV * ZW; // Beta_Zx
Wy        = A_m.t() * b / static_cast<float>(m);
FTF       = F.t() * F;
gamma     = lambda_f*alpha_f*F_fTF_f;
U         = glasso_factor_fast2(WMW,FTF,rho,gamma);
L         = U.t();

/* Ciclo principale  */
for (k=0; k<maxiter; k++) {

  /* Se cambia rho dato il relaxation */
  if(k>0){
      if(rho_store(k-1)!=rho){

          U = glasso_factor_fast2(WMW, FTF, rho,gamma);
          L = U.t();
      }
  }

  /* Aggiornamento stima coefficienti */
  q = WMy+ rho *F.t() * (z - u);              // quantit\`a temporanea
  x = solve(trimatu(U), solve(trimatl(L), q));// aggiornamento regressione x (XTX)XT(Y-Zv)
  v = v_LS - P_ZW * x;                        // aggiornamento regressione v (ZTZ)ZT(Y-XB)
```

```
z_old = z; /* re-definizione per aggiornamento */

/* Aggiornamento z fused LASSO */
arma::vec u_f = u.subvec(0, F_f.n_rows-1);
arma::vec z_f = lasso_prox(F_f * x + u_f, (lambda_f*(1-alpha_f))/ rho);

z.subvec(0, F_f.n_rows-1) = z_f;

g_id_init=F_f.n_rows;

/* Aggiornamento z group LASSO per variabile */
for (int g=0; g<n; g++) {

  g_id_end      = g_id_init +m -1;
  arma::vec u_g = u.subvec(g_id_init, g_id_end);
  arma::mat F_g = F.submat(g_id_init, 0, g_id_end, p-1);
  arma::vec z_g                 = glasso_prox(F_g * x + u_g, lambda_l_T/ (rho*pesi(g)));
  z.subvec(g_id_init, g_id_end) = z_g;
  g_id_init                     = g_id_end +1;
}

g_id_init=F_l.n_rows+F_f.n_rows;

/* Aggiornamento z group LASSO per tempo */
for (int g=0; g<m; g++) {

  g_id_end      = g_id_init +n -1;
  arma::vec u_g = u.subvec(g_id_init, g_id_end);
  arma::mat F_g = F.submat(g_id_init, 0, g_id_end, p-1);
  //cout<<F_g<<endl;
  arma::vec z_g                 = glasso_prox(F_g * x + u_g, lambda_l_p / (rho));
  z.subvec(g_id_init, g_id_end) = z_g;
  g_id_init                     = g_id_end +1;
}


/* Aggiornamento variabile duale */
u = u + F * x - z;

/* Diagnostica */
h_r_norm(k) = norm(glasso_residual(F, x, z), 2);
h_s_norm(k) = norm(glasso_dual_residual(F, z, z_old, rho), 2);
if (norm(F * x, 2) > norm(-z, 2)) {
  h_eps_pri(k) = std::sqrt(z.n_elem) * abstol + reltol * norm(F * x, 2);
} else {
  h_eps_pri(k) = std::sqrt(z.n_elem) * abstol + reltol * norm(-z, 2);
}
h_eps_dual(k) = std::sqrt(z.n_elem) * abstol + reltol * norm(F.t() * rho * u, 2);

/* :::::::::::::::::::::::::::::::::::::::::::::::::::::: */
/* Relaxation                                             */

if (rho_relaxation) {
  if (h_r_norm(k) > alpha * h_s_norm(k)) {
    rho = rho * tau;
```

```
        u   = u / tau;
      } else if (h_s_norm(k) > alpha * h_r_norm(k)) {
        rho = rho / tau;
        u   = u * tau;
      }
    }
    rho_store(k+1) = rho;


    /* Criterio di fermata */

    if ((h_r_norm(k) < h_eps_pri(k)) && (h_s_norm(k) < h_eps_dual(k))) {
      break;
    }
  }

  /* Tempo per la stima */
  elTime = timer.toc();

  /* Risultati       */
  Rcpp::List output;
  output["x"]        = x;
  output["u"]        = u;
  output["z"]        = z;
  output["objval"]   = h_objval;
  output["niter"]    = k;
  output["r_norm"]   = h_r_norm;
  output["s_norm"]   = h_s_norm;
  output["eps_pri"]  = h_eps_pri;
  output["eps_dual"] = h_eps_dual;
  output["F"] = F;
  if ((k+1) < maxiter) {
    output["convergence"] = 1.0;
  } else {
    output["convergence"] = 1.0;
  }
  output["fitted"]=A*x;
  output["D.Matrix"]=A;
  output["eltime"] = elTime;
  output["rho"]    = rho_store;
  output["OLS"]=pesi;
  output["fixed"]=v;

  /* Return output      */
  return(output);
}
```

Listing A.1: *ADMM*

Bibliografia

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 2:255–265.

Azzalini, A. and Scarpa, B. (2012). *Data analysis and data mining*. Oxford University Press.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 160(901):268–282.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122.

Byod, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Diebold, F. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130:337–364.

Fahrmeir, L. and Kaufmann, H. (1991). On kalman filtering, posterior mode estimation and fisher scoring in dynamic exponential family regression. *Metrika*, 38:37–60.

Friedman, J., Hastie, T., and Tibishirani, R. (2010). note on the group lasso and sparse group lasso.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers Mathematics with Applications*, 2(1):17–40.

Glowiski, R. and Marrocco, A. (1975). Sur l'approximation, par elements finis d'ordre un, et la resultion, par penalisation-dualite, d'une classe de problems de dirichlet

non linears. *Revue Francais d'automatique, informatique, et recherche operationelle*, 9:41–76.

Goeman, J. (2008). Autocorrelated logistic ridge regression for prediction based on proteomics spectra. *Statistical Application in genetics and molecular biology*, 10.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hastie, T., Tibishirani, R., and Friedman, J. (2009). *The elements of stastistical learning*. Springer.

Hestens, M. (1969). *Multiplier and gradient methods in Computing method in Optimization problems*. Academic Press.

Hodges, J. (2014). *Richly parameterized linear models, additive, time series, and Spatial Models using random effects*. Chapman Hall/Crc.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.

Hunter, D. and Lange, K. (2004). A tutorial on mm algorithms. *The american statistician*, 58.

Kalman, R. (1960). A new approach to linear filtering and prediction theory. *Journal of basic engineering*, 82:35–45.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized linear models with random effects*. Chapman Hall/Crc.

Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112.

Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal statistical Society*, 14.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. *The Journal of Business*, 60(4):473–489.

Ortega (1970). Iterative solutions of nonlinearequations in several variables. *New York: Academic*, 58:253–255.

Powell, M. (1969). *A method for nonlinear constraints in minimization problems.* Academic Press.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172.

Ripley, D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6:461–464.

Snijders, T. A. B. and Bosker, R. J. (2000). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* Sage, London, Thousand Oaks.

Taylor, J. and Tibishirani, R. (2012). Degrees of freedom in lasso problems. *The annals of statistics*, 40:1198–1232.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Triantafyllopoulos, K. (2021). *Bayesian Inference of State Space Models.* Springer.

van Wieringen, W. (2021). Lectures and notes on ridge regression. *Lectures and notes on Ridge Regression*, 0.40.

Zou, H. (2012). The adaptive lasso and its oracle properties. *Journal of america statistical association*, pages 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.