Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione finale

R-SADA: un algoritmo di campionamento basato su ranghi mediante aumento dei dati e la sua applicazione nel monitoraggio online di flussi di big data

Relatore: Prof. Guido Masarotto
Dipartimento di Scienze Statistiche

Laureando: Monte Valentina
Matricola N. 1232342

Anno Accademico 2021/2022

Indice

Introduzione

In molte applicazioni del controllo della qualità il monitoraggio di un processo coinvolge un gran numero di variabili. Si vuole ottenere informazioni complete al fine di garantire un rapido rilevamento dei cambiamenti del processo che possono eventualmente verificarsi in qualsiasi variabile. Tuttavia, le informazioni complete non sono sempre disponibili durante il controllo online dei flussi di big data a causa delle limitazioni delle risorse.

Questa relazione ha lo scopo di presentare un algoritmo di monitoraggio e campionamento basato su ranghi e sull'aumento dei dati per rilevare rapidamente gli spostamenti medi in un processo quando si ha a disposizione solo una parte limitata delle osservazioni.

Questo metodo è stato introdotto da Xiaochen Xiana, Chen Zhangb, Scott Bonka e Kaibo Liua nell'articolo: *Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation* (2019).

Il lavoro si compone di 5 capitoli.

Nel primo verrà introdotto l'ambito in cui si inserisce l'algoritmo, il controllo statistico della qualità, con degli accenni teorici agli strumenti che verranno utilizzati nei capitoli successivi.

Il secondo capitolo svilupperà il metodo R-SADA con attenzione sugli aspetti concettuali, la costruzione della carta di controllo e le sue proprietà.

Nel terzo capitolo si mostrerà una possibile implementazione dell'algoritmo in R con una spiegazione dettagliata delle funzioni che entrano in gioco.

Il quarto capitolo presenta un'applicazione a dati simulati con relativo calcolo della carta di controllo e calcolo dei limiti (dinamici e con metodo bootstrap).

L'ultimo capitolo invece mette in luce le criticità presenti nell'algoritmo raccontato e una possibile soluzione con corrispondente implementazione.

Capitolo 1

Controllo statistico della qualità

In questo capitolo verrà contestualizzato il macro argomento sul quale è basata l'intera relazione prestando particolare interesse alle nozioni basilari che verranno richiamate in seguito in contesti più complessi.

1.1 Introduzione

Il controllo e il miglioramento della qualità sono diventati un'importante strategia aziendale per molte organizzazioni: produttori, distributori, società di trasporto, organizzazioni di servizi finanziari, fornitori di assistenza sanitaria e agenzie governative. Mantenere un livello elevato di qualità del prodotto o del servizio infatti fornisce un vantaggio competitivo.

Si può definire il Controllo Statistico della Qualità (o Statistical Process Control, in seguito SPC) come una metodologia che, in riferimento ad una determinata attività, operazione, fase o processo caratterizzato da ripetitività, fa ricorso a tecniche statistiche al fine di definire, analizzare e verificare le condizioni che determinano la variabilità dell'oggetto di analisi.

In modo più sintetico, facendo riferimento alla definizione fornita da Juran si può definire l'SPC come "l'applicazione di tecniche statistiche per comprendere ed analizzare le variabilità di un processo".

Ogni processo racchiude al suo interno una variabilità che non può essere eliminata originata da una serie di fluttuazioni interne al processo, risultato di numerose piccole cause che operano casualmente (dette cause comuni o ca-

suali). Sulla variabilità del processo possono però intervenire fattori esterni che ne alterano la variabilità naturale e generano una variabilità non prevedibile che disturba il funzionamento del processo. Tali fattori sono denominati cause speciali di variazione che possono indicare il verificarsi di problemi oppure, al contrario, l'insorgere di novità interessanti da esplorare. Quando questo accade il processo viene definito "fuori controllo" (o out control), in seguito "OC".

1.2 Carte di controllo

Al fine di ottenere livelli di qualità accettabili diventa determinante intraprendere un'azione di monitoraggio della variabilità del processo produttivo.

Le carte di monitoraggio servono per individuare gli allontanamenti dalla condizione definita "in controllo" per cercare di attenuare i possibili danni.

Le carte di controllo si distinguono in due gruppi:

- carte di controllo per la Fase 1: viene definito un periodo e si analizzano i dati storici riferiti a questo; sono retrospettive e servono per capire se il processo è stato in controllo durante un periodo predefinito; servono per capire le modalità con cui un processo va fuori controllo;
- carte di controllo per la Fase 2: i dati vengono analizzati sequenzialmente man mano che vengono rilevati con lo scopo di dare un allarme non appena si rilevino spostamenti dallo stato di stabilità; sono carte di tipo prospettico.

In generale le carte di monitoraggio sorvegliano un processo, oppure una o più sue caratteristiche, utilizzando una statistica di controllo calcolata al tempo t: Wt. Se tale statistica fuoriesce da dei predefiniti limiti di controllo, allora al tempo t la carta segnala un allarme.

Si tratta essenzialmente di rappresentazioni grafiche di un processo nel tempo che, basandosi su teorie statistiche, rimangono di facile interpretazione e utilizzo anche per utenti meno esperti.

In letteratura esistono diversi tipi di carte di controllo ma la forma generale è quella riportata in figura.

Le tre linee orizzontali continue chiamate linea centrale (CL), limite superiore di controllo (UCL) e limite inferiore di controllo (LCL) definiscono la tendenza centrale e un range di variazione naturale per i valori riportati sul grafico. I limiti inferiori e superiori sono calcolati in base a una distribuzione di frequenza teorica che cambia in funzione del tipo di dati che vengono analizzati (gaussiana, Poisson, binomiale, ...).

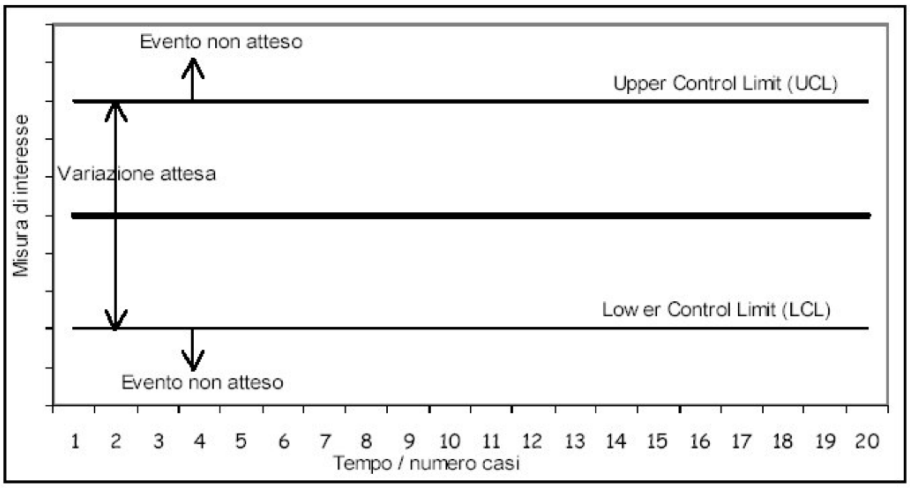

Figura 1.1: Carta di Controllo

Una volta definiti i limiti di controllo, plottando i dati all'interno del grafico, la carta consente di individuare eventuali andamenti sistematici (pattern) dei valori che rappresentano il processo nel tempo e di stabilire se ciascun punto cade all'interno o all'esterno dei limiti imposti. In questo modo si individua immediatamente un processo fuori controllo.

1.2.1 Average Run Length e Detection Delay

L'Average Run Length , in seguito ARL, non è altro che il valore atteso della Run Length (RL) così definita:

$$RL = min\{t : W_t \notin [LCL_t, UCL_t]\} \tag{1.1}$$

la quale indica il numero di istanti di tempo tra l'inizio della sorveglianza e il primo allarme.

Risulta molto utile perchè indica quanto impiega una carta di controllo a segnalare un'allarme (sia questo corretto o meno) e la distribuzione della RL (o dell'ARL) può essere utilizzata per valutarne la performance.

Legato all'ARL è anche il Detection Delay, DD nel seguito, ovvero gli istanti di tempo che intercorrono da quando il processo va fuori controllo a quando viene chiamato l'allarme per segnalarlo.

1.3 Un esempio di carta di controllo: CUSUM

Si procede ora con il definire nel dettaglio una particolare carta di controllo che verrà ripresa più volte nei capitoli successivi.

La carta CUSUM (Cumulative Sum) è una valida opzione per individuare piccoli shift nei parametri soprattutto quando la numerosità campionaria risulta ridotta in cui si incorre nel rischio di generare intervalli di accettazione più ampi e, di conseguenza, una maggiore probabilità di accettare l'ipotesi nulla quando è falsa .

La carta CUSUM risolve questo problema in quanto incorpora tutta l'informazione della sequenza dei campioni estratti basandosi sull'idea di sommare gli scostamenti (positivi o negativi) dal valore centrale e quindi risulta più sensibile ad un aumento o ad una diminuzione della caratteristica che si sta monitorando.

Formalizzazione

Siano $y_1, ..., y_m$ determinazioni indipendenti e identicamente distribuite di una variabile casuale, eventualmente multivariata, con funzione di densità $f(.)$. Si consideri il sistema di ipotesi

$$\begin{cases} H_0 : & f(.) = f_0(.) \\ H_1 : & f(.) = f_1(.) \end{cases} \tag{1.2}$$

Allora, per ogni L, il test che rifiuta H_0 quando

$$W = \sum_{i=1}^{m} \log \frac{f_1(y_i)}{f_0(y_i)} > L \tag{1.3}$$

è il test più potente tra tutti quelli che hanno un errore di primo tipo uguale o inferiore a

$$\alpha = PR_{H_0}(W > L)$$

Sia $y_1, y_2, ...$ una successione di variabili casuali indipendenti (eventualmente multivariate) tali che

$$y_t \sim \begin{cases} f_0(.) & t < \tau \\ f_1(.) & t \geq \tau \end{cases} \tag{1.4}$$

dove τ è non noto mentre $f_0(.)$ e $f_1(.)$ sono due densità completamente note. Allora la carta di controllo che segnala un allarme quando $W_t > L$ dove

$$W_t = \begin{cases} 0 & t = 0 \\ max\left(0, W_{t-1} + \log \dfrac{f_1(y_t)}{f_0(y_t)}\right) & t > 0 \end{cases} \tag{1.5}$$

e L è una costante positiva, minimizza il ritardo atteso

$$\lim_{\tau \to +\infty} E_{f_1}(RL - \tau + 1 | RL \geq \tau) \tag{1.6}$$

tra tutte le carte di controllo con ARL in controllo uguale o maggiore a quella dello schema descritto.

Capitolo 2

R-SADA

In questo capitolo verrà introdotto il contesto nel quale viene applicata la carta di controllo R-SADA proposta da Xiaochen Xian et al. (2021). Si cercherà di capire in cosa si differenzia da quelle già esistenti (Top-r Adaptive Sampling (TRAS), Nonparametric Anti-rank based Sampling (NAS) e Top-r Cusum) nella risoluzione del problema dei dati incompleti, verranno evidenziate le migliorie introdotte e saranno infine definite le quantità coinvolte.

2.1 Carte di controllo e dati mancanti

L'avanzamento della tecnologia dei sensori per il monitoraggio di processi ha reso sempre più disponibili flussi di big data. Per big data in questo contesto si intendono flussi di dati di grandi dimensioni di serie multiple di osservazioni in tempo reale, continue e ordinate in sequenza.

Il monitoraggio online di questi flussi di dati è ancora complesso a causa del fatto che gli eventi anomali che caratterizzano un processo qualsiasi sono sconosciuti in anticipo e i flussi di big data hanno un elevato volume, dimensionalità, richiedono un elevato spazio di archiviazione, potenza di calcolo e tempo di elaborazione. Spesso si arriva quindi a prendere delle decisioni sulla base di osservazioni parziali.

Le cause che portano ad avere rilevamenti parziali si possono riassumere nelle seguenti:

1. limitazione del numero di sensori, si possono ottenere misurazioni solo di un sottoinsieme delle variabili interessate ad ogni momento di acquisizione dei dati;

2. limitazione della durata della batteria dei sensori, solo un sottoinsieme di sensori può essere attivato per raccogliere dati in tempo reale;

3. limitazione della larghezza di banda di comunicazione, solo un sottoinsieme delle misurazioni possono essere ritrasmesse al data center per l'elaborazione in tempo reale.

Malgrado questa consapevolezza la letteratura tradizionale nell'ambito del controllo statistico della qualità si basa sul presupposto che tutte le osservazioni di tutte le variabili siano completamente accessibili in tempo reale.

2.2 Esempi di ambiti applicativi

Nell'ambito ambientale spesso si ricorre alla raccolta delle misurazioni con più sensori disposti in varie località al fine di prevenire ed evitare eventi catastrofici quali terremoti, incendi, tornado e smottamenti. Per rilevare in modo rapido cercando di attenuare così le perdite ambientali ed economiche si cerca di accedere alle osservazioni complete di ciascuna località al momento dell'acquisizione dei dati. Spesso però questo non è possibile a causa della durata limitata delle batterie dei sensori e dell'elevato costo da sostenere per la loro sostituzione; in queste condizioni si ha solamente un sottoinsieme dei sensori che sono attivi per la sorveglianza e quindi solo le osservazioni parziali sono effettivamente consultabili in tempo reale.

In molte altre applicazioni anche la larghezza di banda di comunicazione dei dati può impedire l'acquisizione o l'accesso a osservazioni complete in tempo reale. Questo avviene per esempio durante il monitoraggio del verificarsi di brillamenti solari in cui solo le informazioni parziali dell'immagine catturate dal satellite possono essere trasmesse sulla terra per l'analisi in tempo reale a causa del vincolo della larghezza di banda di comunicazione, sebbene sia possibile registrare le informazioni complete sull'immagine e disponibile per l'analisi offline.

In entrambi gli esempi precedenti, solo una parte dei flussi di dati e di variabili è osservabile a causa di vincoli di risorse, mentre la modificazione dello stato in controllo del processo in esame può verificarsi a causa di qualsiasi variabile dell'intero flusso di dati.

2.3 Formalizzazione matematica del problema

Si suppone di monitorare p variabili indipendenti e identicamente distribuite (i.i.d.) e le misurazioni di ciascuna variabile, denotate come $X_j(t)$ $j \in P = (1, 2, \ldots, p)$, formano un flusso di dati nel tempo di osservazione $t = 1, 2, \ldots$.

L'assunzione i.i.d. dei flussi di dati è stata utilizzata nell'impostazione della letteratura di monitoraggio ad alta dimensionalità non solo per semplicità ma anche perché può essere considerata soddisfacente quando gli $X_j(t)$ sono selezionati come residui di alcuni modelli spazio-temporali.

Con $X(t) = (X_1(t), X_2(t), ..., X_p(t))'$ indichiamo le misurazioni delle p variabili al tempo t: a causa di vincoli di risorse, solo q ($q < p$) delle variabili p sono "osservabili" per ogni t, dove q è limitato dalla disponibilità pratica di risorse di monitoraggio e quindi predeterminato dal contesto applicativo. L'insieme delle variabili osservabili al tempo t è indicato come $O(t)$, e i dati osservati sono indicati come $X^O(t)$.

Quando il processo è in controllo (IC), si presume che $X(t)$ sia i.i.d. in diversi punti temporali e ogni flusso di dati abbia una funzione di probabilità cumulativa $F(x)$ e una funzione di densità $f(x)$. Le funzioni di distribuzione $F(x)$ e $f(x)$ possono essere acquisite empiricamente sulla base dell'analisi dei dati storici IC di Fase I. Questa relazione, come l'artiolo originale, si concentra invece sul monitoraggio di Fase II.

Senza perdita di generalità, si presume che $X(t)$ abbia una media IC $\mu = \mathbf{0} = (0, 0, ..., 0)'$ e ogni variabile abbia una deviazione standard IC di 1 (risultato di centramento e standardizzazione dei dati grezzi). Il processo diventa fuori controllo (OC) se almeno una delle variabili ha uno spostamento medio in un determinato istante sconosciuto τ, tale che il vettore medio di $X(t)$ cambia da $\mu = \mathbf{0}$ a $\mu \neq \mathbf{0}$.

Per evidenziare l' idea principale del metodo proposto, di seguito ci concentreremo sulla rilevazione degli spostamenti medi verso l'alto.

Sulla base della formulazione sopra riportata l'obiettivo principale è trovare una metodologia per decidere in modo intelligente e sequenziale le variabili più informative da osservare in ogni momento al fine di rilevare rapidamente lo spostamento medio del processo mantenendo un ARL in controllo specificato a priori.

Nel monitoraggio di flussi di big data con osservazioni parziali, lo stato del sistema dipende sia dalle variabili osservate che da quelle non osservate. Si propone di sfruttare l'aumento dei dati per aumentare analiticamente le statistiche delle variabili non osservate per facilitare il monitoraggio e il campionamento di tutte le variabili coinvolte nel processo. L'aumento dei dati si riferisce a un tipo di metodi che consiste nell'aggiunta di informazioni o variabili latenti ai dati originali "non osservabili" o "mancanti", in modo tale che il problema diventi trattabile.

L'idea risulta concettualmente valida ma nell'ambito applicativo gli ostacoli che si incontrano sono molteplici:

- non sempre si riesce a trovare le relazioni che legano le variabili osservabili con quelle non osservabili per integrare efficacemente i metodi di aumento con lo schema di monitoraggio e campionamento;
- visto che le risorse sono limitate, sono sconosciuti il punto di cambiamento di stato del sistema e il numero di variabili che ne sono la causa, può succedere che solo alcune variabili fuori controllo vengano osservate dopo che lo spostamento si è verificato nel sistema. I valori aumentati servono proprio per rendere più facile il rilevamento di questo cambiamento anche quando viene osservato un numero molto piccolo di variabili fuori controllo;
- i flussi di big data potrebbero non seguire alcune distribuzioni ben note. Di conseguenza, si desidera che il metodo di aumento non sia limitato ad una certa distribuzione.

2.4 R-SADA

2.4.1 Introduzione

Per affrontare le problematiche sopra riportate, viene proposto dagli autori un metodo di monitoraggio e campionamento CUSUM basato sui ranghi, denominato *Rank-based Sampling Algorithm by Data Augmentation* (R-SADA), per rilevare più velocemente i cambiamenti di processo nel contesto di osservazioni parziali aumentando i dati non osservabili utilizzando le misurazioni di quelli osservati.

Il metodo di monitoraggio e campionamento proposto è parametrico poiché incorpora la funzione di ripartizione e di densità della distribuzione del processo nella fase di aumento dinamico dei dati anche se si basa sui ranghi.

Le motivazioni che spingono ad utilizzare un metodo basato sui ranghi sono molteplici:

- ha la caratteristica di avere un robusto potere di rilevamento per le distribuzioni generali;
- le informazioni sul rango dei flussi di dati sono naturalmente dipendenti, il che consente di aumentare dinamicamente le statistiche utili relative allo stato del sistema in base ai valori osservati;
- poiché l'informazione di rango di una variabile è direttamente associata a tutte le altre variabili, il metodo proposto è sensibile per attivare rapidamente un allarme anche se viene osservato solo un numero ridotto di variabili OC;

- può essere efficacemente integrato con lo schema di monitoraggio e campionamento assicurando due proprietà interessanti: nessuna variabile sarà lasciata inosservata per lungo tempo quando il processo è IC, quando il processo diventa OC il metodo tende a continuare a osservare la sospetta variabile OC portando ad una rapida individuazione delle cause assegnabili.

Si vuole in primo luogo esaminare la letteratura esistente relativa ad altri approcci altamente correlati con R-SADA: gli approcci SPC basati su ranghi, approcci SPC multivariati e SPC con osservazioni parziali.

Cenni procedure SPC basate su ranghi e carte di controllo multivariate

I metodi basati sui ranghi sono utilizzati quasi esclusivamente in SPC non parametrici. L'idea principale dei metodi basati sui ranghi è di utilizzare il rango tra le osservazioni invece delle osservazioni stesse pertanto non è necessario alcun modello parametrico.

Le metodologie basate sui ranghi possono essere suddivise in due categorie: per i casi univariati e per quelli multivariati.

Per riportare un esempio che tornerà utile in seguito Qiu e Hawkins (2001, 2003) hanno proposto una carta di controllo CUSUM multivariata non parametrica basata sul monitoraggio dell'anti-rango delle variabili. In particolare, il vettore anti-rango di $X(t) = (X_1(t), X_2(t), ..., X_p(t))$, indicato come $B(t) = (B_1(t), B_2(t), ..., B_p(t))$, è una permutazione di $(1, 2, \ldots, p)'$, tale che $X_{B_1(t)}(t) \leq X_{B_2(t)}(t) \leq ... \leq X_{B_p(t)}(t)$. Una statistica $\zeta(t) = (\zeta_1(t), \zeta_2(t), ..., \zeta_p(t))$ può essere costruita sulla base dell'ultimo $B_p(t)$ anti-rango, dove $\zeta_j(t) = \mathbb{I}(B_p(t) = j)$.

Gli autori hanno dimostrato che rilevando i cambiamenti nelle distribuzioni di $X(t)$ (con ipotesi nulla $H_0 : \mu_1 = ... = \mu_p$) equivale a rilevare i cambiamenti nell'aspettativa IC di $\zeta(t)$, cioè $\mathbb{E}[\zeta(t)] = (g_1, g_2, ..., g_p)$ dove g_i è la probabilità che la i-esima variabile assuma il valore più grande tra tutte le misurazioni sotto H_0. La statistica di monitoraggio descrive dunque lo stato del processo in base alla differenza tra l'anti-rango osservato $\zeta(t)$ e la sua aspettativa IC.

Per monitorare più variabili contemporaneamente e sfruttare le informazioni globali, sono stati sviluppati molti metodi che partendo da singole statistiche locali danno luogo a statistiche globali.

Viene calcolata in primo luogo una statistica locale per ciascuna variabile mediante alcune carte di controllo univariate (ad es. CUSUM) e poi vengono combinate le statistiche in modo aggregato per il monitoraggio globale.

Ad esempio, Woodall e Ncube (1985) [6] e Tartakovsky et al. (2006) hanno suggerito di utilizzare le più grandi statistiche CUSUM locali come statistica di monitoraggio globale, mentre Mei (2010) ha introdotto l'utilizzo della somma di tutte le statistiche locali. Questi due schemi sono indicati con T_{max} e T_{sum}.

Mei (2011) ha inoltre proposto di sommare le più grandi statistiche locali (denominate Top-r) come compromesso tra T_{max} e T_{sum}.

Le prestazioni delle statistiche dipendono tuttavia dal numero di variabili OC e se tutte le variabili OC non possono essere osservate durante il monitoraggio del flusso di dati si rischia un ritardo nel rilevamento. Di conseguenza, è necessario uno schema di monitoraggio migliore per attivare rapidamente un allarme, che non dipenda dal numero di variabili OC osservate per volta.

Cenni SPC con osservazioni parziali

Gli studi riguardo SPC con osservazioni parziali possono essere classificati in due tipologie:

- la prima categoria si concentra sul campionamento di osservazioni parziali nel dominio temporale, ovvero sulla regolazione adattiva degli intervalli di campionamento in base ai valori osservati. Questa linea di ricerca è nota come carta di controllo dell'intervallo di campionamento variabile (VSI). La strategia non risulta adatta al problema presentato in quanto richiede osservazioni complete di tutti i flussi di dati a ogni tempo di campionamento;

- la seconda categoria considera l'SPC con strategie di campionamento spaziale, ovvero osservando solo una parte delle variabili e determinando in modo adattivo quali variabili osservare in ogni momento. In questa categoria, ci sono due diversi approcci per gestire i flussi di dati non osservabili:

  - il primo approccio consiste nell'utilizzare solo le variabili osservabili per costruire statistiche di monitoraggio. Tuttavia non è adatto per grandi flussi di dati a causa del suo elevato carico computazionale e necessita di una struttura di rete di tipo bayesiano;
  - il secondo approccio è l'aumento dei dati.

L'aumento dei dati proprio come suggerisce il termine indica l'aggiunta di dati a disposizione partendo dai dati osservati. È stato ampiamente applicato nell'analisi statistica e nelle applicazioni di apprendimento automatico come la stima dei parametri, la progettazione di esperimenti, la regressione, le simulazioni Monte Carlo e costruzioni di dati di addestramento.

La strategia dell'aumento dei dati sembra essere la più promettente per il quesito proposto e nella letteratura esistono già dei metodi che ne fanno uso:

- Top-r Adaptive Sampling (TRAS): si concentra sul monitoraggio dei flussi di dati di grandi dimensioni che seguono le distribuzioni normali, si basa su un approccio Top-r CUSUM e introduce un parametro di imputazione costante alla statistica locale delle variabili non osservabili che compensa le osservazioni non effettuate;
- Nonparametric Anti-rank based Sampling (NAS): generalizza l'idea riportata sopra ma viene utilizzato per monitorare flussi di dati distribuiti arbitrariamente incorporando il parametro di imputazione con un metodo basato sui ranghi.

I metodi sopra proposti considerano il parametro di imputazione predeterminato e costante pertanto la stima dei dati non osservati non è informativa e non si adatta alle misurazioni online perchè può introdurre distorsioni o incertezze nello schema di monitoraggio ed in qualche caso degradare le prestazioni di rilevamento.

2.4.2 Costruzione del vettore aumentato di osservazioni basato su anti-rank

Si ricorda che $X(t)$ è il vettore che conterrà i valori derivanti dalla misurazione di p variabili al tempo t, ed $X^O(t)$ è il sottoinsieme osservabile di $X(t)$. Nello scenario di osservazioni parziali si vuole costruire un vettore dinamico aumentato basato sui dati osservabili $X^O(t)$ il quale sarà poi utilizzato per la costruzione di un CUSUM multivariato.

L'approccio proposto si ispira alla procedura anti-rango brevemente esposta sopra ma in questo contesto l'indicatore anti-rango $\zeta(t)$ viene generalizzato. La differenza principale rispetto ai metodi proposti precedentemente è che si mira a costruire dinamicamente e analiticamente il vettore aumentato senza utilizzare un valore costante e deterministico per definire l'aumento.

Viene definito il vettore aumentato come:

$$\eta(t) = \mathbb{E}[\zeta(t)|X^O(t)] \tag{2.1}$$

Pertanto per ogni variabile j appartenente a P

$$\eta_j(t) = \mathbb{E}[\zeta_j(t)|X^O(t)] = \mathbb{P}(\zeta_j(t) = 1|X^O(t)) \tag{2.2}$$

$\eta(t)$ rappresenta la probabilità che una variabile sia la più grande tra tutte le p variabili. Quindi verificare la consistenza di $\eta(t)$ e la sua media IC

$$\mathbb{E}[\eta(t)] = \mathbb{E}[\mathbb{E}[\zeta(t)|X^O(t)]] = \mathbb{E}[\zeta(t)] = g \tag{2.3}$$

equivale a verificare

$$H_0 : \mu_1 = \mu_2 = ... = \mu_p \tag{2.4}$$

Il calcolo e l'aumento di $\eta_j(t)$ basato sull'equazione 2.2 si basano sul vero vettore medio di $X(t)$, cioè, $\mu(t)$.

Per calcolare $\eta(t)$ si suppone ora che solamente una variabile abbia uno spostamento medio nello scenario OC.

$$H_1 : \mu_{j_{OC}} = \mu_{OC} > \mu_1 = ... = \mu_{j_{OC-1}} = \mu_{j_{OC+1}} = ... = \mu_p = 0 \tag{2.5}$$

dove $j_{OC} \in P$ è l'unica variabile OC. Tale ipotesi risulta ragionavole in quanto le variabili che subiscono uno spostamento sono solitamente sparse nell'ambito dei big data. L'entità dello spostamento OC μ_{OC} è sconosciuta in anticipo. Si utilizza un parametro $\mu_{min} > 0$ per rappresentare la grandezza più piccola interessante degli spostamenti medi da rilevare.

Si suppone $i(t) = argmax_{j \in O(t)} X_j(t)$ sia l'indice della più grande variabile osservabile al tempo t. Per costruire un vettore dinamico aumentato che reagisce tempestivamente a spostamenti sospetti la probabilità viene aggiornata in base a tre diversi scenari:

1. se $j = i(t)$

$$\eta_j(t) = \frac{F(X_{i(t)}(t))^{p-q} \sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t))}{\sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t)) + (p-q)} + \frac{F(X_{i(t)}(t))^{p-q-1} F(X_{i(t)} - \mu_{min})(p-q)}{\sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t)) + (p-q)} \tag{2.6}$$

2. se $j \in O(t)$ e $j \neq i(t)$

$$\eta_j(t) = 0 \tag{2.7}$$

ovvero quando una variabile osservata j non è tra le più grandi in $O(t)$, $\eta_j(t)$ sarà impostato a zero poiché non può essere la più grande tra tutte le p variabili.

3. se $j \notin O(t)$

$$\eta_j(t) = \frac{(1 - F(X_{i(t)}(t))^{p-q} \sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t))}{(p-q)(\sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t)) + (p-q))} + \frac{1 - F(X_{i(t)}(t))^{p-q-1} F(X_{i(t)} - \mu_{min})(p-q)}{\sum_{l \in O(t)} f(X_l(t) - \mu_{min})/f(X_l(t)) + (p-q)} \tag{2.8}$$

Si noti che $\sum_{j=1}^{p} \eta_j(t) = 1$.

Il valore $\eta_j(t)$ aumentato cambia dinamicamente in funzione dei valori osservati $X^O(t)$ anziché essere una costante; ci si aspetta pertanto che questo vettore dinamico aumentato riconosca meglio lo stato del sistema in tempo reale e reagisca tempestivamente a cambiamenti sospetti.

Ad esempio, dall'ultima equazione possiamo vedere che quando $X_{i(t)}(t)$ è grande, il valore di $\eta_j(t)$ ($j \notin O(t)$) diventa più piccolo. Ciò significa che la probabilità che una variabile non osservata sia la più grande diminuisce. Inoltre, il valore aumentato $\eta_j(t)$ si basa su tutte le osservazioni $X^O(t)$ per sfruttare le informazioni disponibili nella misura massima possibile. Questa procedura è quindi fondamentalmente diversa dagli schemi di monitoraggio convenzionali come T_{max} e T_{sum}, in cui le singole statistiche sono statistiche "locali", nel senso che ciascuna di esse si basa esclusivamente sul flusso di dati corrispondente.

Vantaggi del metodo proposto

Una volta individuate le caratteristiche che contraddistinguono il metodo proposto rispetto a quelli precedenti se ne presentano i vantaggi:

- consente di incorporare modelli parametrici in un framework basato su ranghi, ed eredita così la flessibilità per gestire vari tipi di distribuzioni;
- il vettore aumentato dinamico dipende dai flussi di dati osservabili pertanto lo spostamento che si verifica in corrispondenza di qualsiasi variabile osservabile avrà un'influenza sulla distribuzione dell'intero vettore aumentato, il che sarà molto vantaggioso per il rapido rilevamento di cambiamenti;
- evita la sfida di scegliere una combinazione efficace di statistiche locali per il monitoraggio globale e il bilanciamento tra T_{max} e T_{sum}.

2.4.3 Carta di controllo R-SADA e algoritmo di campionamento

Ad ogni momento di acquisizione, il vettore dinamico aumentato $\eta(t)$ viene costruito sulla base delle osservazioni parziali disponibili andando poi a contribuire ad una statistica di monitoraggio globale. Se la statistica di monitoraggio supera un limite di controllo (ARL IC pre-specificato), la carta di controllo R-SADA attiva un allarme e interrompe il processo.

Statistica di monitoraggio

Bisogna verificare il valore atteso $\mathbb{E}[\eta(t)] = g$ riguardo l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = ... = \mu_p$. Per farlo si decide di adottare le statistiche di monitoraggio dell'approccio CUSUM proposte da Qiu e Hawkins (2001). Siano $S_t^{(1)}$ e $S_t^{(2)}$ statistiche CUSUM per $\eta(t)$ e g rispettivamente. Allora:

$$\begin{cases} S_t^{(1)} = g, & S_t^{(2)} = g & \text{se } C_t <= k, \\ S_t^{(1)} = \left(S_{t-1}^{(1)} + \eta(t)\right)(C_t - k)/C_t & S_t^{(2)} = \left(S_{t-1}^{(1)} + g\right)(C_t - k)/C_t & \text{se } Ct > k \end{cases} \tag{2.9}$$

in cui

$$C_t = \left(S_{t-1}^{(1)} - S_{t-1}^{(2)} + \eta(t) - g\right)' \cdot$$

$$\cdot \text{diag}\left((S_{1,t-1}^{(2)} + g_1)^{-1}, ..., (S_{p,t-1}^{(2)} + g_p)^{-1}\right)\left(S_{t-1}^{(1)} - S_{t-1}^{(2)} + \eta(t) - g\right) \tag{2.10}$$

dove $S_0^{(1)} = S_0^{(2)} = 0$ e k è una costante.

C_t è uno scalare che rappresenta la distanza tra $S_t^{(1)}$ e $S_t^{(2)}$. La quantità k è la tolleranza della CUSUM tale che $S_t^{(1)}$ e $S_t^{(2)}$ siano entrambe in controllo e pari al valore atteso g se la distanza è inferiore a k. Per tenere in considerazione tutti gli elementi la statistica di monitoraggio prende la forma di:

$$y_t = \left(S_t^{(1)} - S_t^{(2)}\right)' \text{diag}\left((S_{1,t}^{(2)})^{-1}, ..., (S_{p,t}^{(2)})^{-1}\right)\left(S_t^{(1)} - S_t^{(2)}\right) \tag{2.11}$$

che è una classica statistica χ^2 di Pearson che misura la differenza statistica tra $S_t^{(1)}$ e $S_t^{(2)}$ quando $k = 0$.

Il processo viene definito fuori controllo se la statistica supera una certa soglia h che dipende dall'ARL in controllo desiderata durante la fase I.

Strategia di campionamento

Per garantire l'efficienza del metodo R-SADA è necessario definire la strategia di campionamento con l'obiettivo di individuare in modo efficace le variabili più sospette di rendere fuori controllo il processo.

Se si osserva uno spostamento verso l'alto che porta una variabile ad essere fuori controllo questa ha una maggiore probabilità di essere il massimo tra tutte le variabili e quindi il suo score η_j nel vettore aumentato sarà maggiore.

Lo stesso concetto vale anche per le variabili OC non osservabili. Allora, dato che $S_t^{(1)}$ è la statistica CUSUM per il vettore aumentato, le variabili associate agli elementi grandi di $S_t^{(1)}$ sono quelle a cui bisogna prestare più attenzione. In modo più preciso viene definito $j_{(l),t}$ l'indice corrispondente alla variabile contenuta nel vettore ordinato in modo decrescente di statistiche $\left(S_{1,t}^{(1)}, ..., S_{p,t}^{(1)}\right)$, in cui $S_{j_{(1),t},t}^{(1)} \geq S_{j_{(2),t},t}^{(1)} \geq ... \geq S_{j_{(p),t},t}^{(1)}$. Al tempo $t+1$ vengono pertanto osservate le variabili i cui indici sono contenuti in

$$O(t+1) = \left\{j_{(1),t}, ..., j_{(q),t}\right\}$$

2.4.4 Proprietà della Carta di controllo R-SADA

In questa sezione, si analizzano due importanti proprietà della carta di controllo R-SADA nelle condizioni in cui il sistema è IC e OC, rispettivamente.

La proprietà IC

Sia U l'insieme di variabili $i \in P$ che non possono mai essere osservate dopo un tempo finito t_0, cioè esiste un tempo t_0 tale che $U = \bigcap_{t=t_0}^{+\infty} P/O(t)$, dove $P/O(t)$ rappresenta il complementare di $O(t)$ rispetto a P. Se $h \to \infty, \mathbb{P}(U = \varnothing) \to 1$ dove $\varnothing$ rappresenta l'insieme vuoto.

La proprietà IC indica che la carta di controllo R-SADA mantiene tutte le variabili sotto sorveglianza quando il processo è IC. In altre parole, non importa quale variabile diventi OC, non sarà trascurata dallo schema proposto sebbene sia possibile osservare solo un numero limitato di variabili alla volta. In particolare, il metodo R-SADA sospetta le variabili non osservabili quando non ci sono prove evidenti che quelle osservate siano OC. In altre parole, tende a campionare le variabili che non sono state monitorate per un po' di tempo.

La proprietà OC

Supponiamo che $k = 0$ e dopo il tempo t_0, vi sia uno spostamento tale che $\mathbb{E}(\eta(t)) \neq g$. Sia j^* la variabile OC con lo spostamento medio maggiore tale che $\mathbb{E}(F(X_{j^*}(t)^{p-q-1})F(X_{j^*}(t) - \mu_{min})) > \frac{p}{q}$.

Allora una volta osservata la variabile j^* all'istante t, c'è una probabilità maggiore di zero che questa variabile j^* venga mantenuta osservata per sempre, cioè $j^* \in O(\tau)$ per ogni $\tau \geq t$.

La proprietà OC mostra che almeno una delle variabili OC con un grande spostamento sarà sempre osservata con una probabilità diversa da zero una volta individuata.

La combinazione delle due proprietà indica che quando il processo è OC, la strategia di campionamento della carta di controllo R-SADA cercherà prima le variabili sospette tra tutte le variabili, quindi si atterrà automaticamente al monitoraggio delle variabili OC sospette e con alta probabilità OC.

Capitolo 3

Metodo R-SADA implementazione in R

In questo capitolo si vuole illustrare passo dopo passo come applicare il metodo proposto ad un insieme di dati pertanto verranno contestualizzate e spiegate le funzioni necessarie per l'applicazione della carta di controllo mediante l'utilizzo del software R.

3.1 Inizializzazione della Carta di Controllo

In primo luogo deve essere inizializzata la carta di controllo. Di seguito viene riportato il codice che lo permette.

```
1 newRSADA <- function(p, q, mu_min, k, F=pnorm, f=dnorm) {
2     list(p=p, q=q, mu=mu_min, k=k, g=rep(1/p,p), F=F, f=f,
3          S1=numeric(p), S2=numeric(p), y=0, eta=NA,
4          O=1:q)
5 }
```

Listing 3.1: Inizializzazione carta di controllo R-SADA

newRSADA riceve come argomenti:

- p, numero totale totale variabili;
- q, numero variabili che possono essere osservate ad ogni istante t;

- mu_min, rappresenta il più piccolo spostamento interessante da rilevare;
- k, costante positiva usata per il CUSUM multivariato;
- F, funzione di ripartizione dalla quale sono stati generati i dati;
- f, funzione di densità dalla quale sono stati generati i dati.

La funzione restituisce una lista con un insieme di oggetti che verranno in seguito utilizzati per calcolare la statistica di controllo riportata nella formula 2.11.

Più precisamente la lista contiene:

- p, q, mu_min, k, F, f che corrispondono agli argomenti passati in input;
- g, valore atteso del vettore aumentato al tempo t=0. Essendo che le variabili sono indipendenti e identicamente distribuite si tratta di un vettore contentente elementi pari a 1/p;
- $S1, S2$, elementi che serviranno per calcolare la statistica di controllo. Inizialmente posti pari a 0 perchè riferiti al tempo t=0;
- y, vettore che conterrà la statistica test;
- η, conterrà il vettore aumentato;
- O, contiene le q variabili da campionare al prossimo istante di tempo tra le p possibili. In questo caso vengono assegnate in modo arbitrario le prime q variabili. O in seguito verrà aggiornato e conterrà le variabili che hanno il corrispondente $S1$ più grande.

3.1.1 Il parametro μ_{min} nel metodo R-SADA

μ_{min} è un parametro unico nel metodo proposto che rappresenta lo spostamento medio più piccolo di interesse, in sostituzione del vero spostamento medio sconosciuto μ_{OC}. Al crescere di μ_{min} (maggiori spostamenti medi) corrispondono cambiamenti più significativi nelle statistiche di monitoraggio e quindi questo viene tradotto in un rilevamento più rapido dello stato OC.

Si può osservare che un μ_{min} più piccolo funziona meglio per spostamenti medi più piccoli e un μ_{min} grande aiuta a rilevare più rapidamente spostamenti medi più grandi. Un valore elevato di μ_{min} aiuta a trovare le variabili OC più rapidamente poiché consente di riallocare le risorse più frequentemente alle variabili non osservabili.

Per piccoli spostamenti, è necessario osservare le variabili OC per un tempo più lungo prima di dare l'allarme. Pertanto un valore inferiore di μ_{min} consente di riallocare le risorse meno frequentemente concentrandosi così maggiormente sulle variabili sospette. Si dovrebbe scegliere opportunamente il valore di μ_{min} sulla base della conoscenza preliminare degli spostamenti e del contesto applicativo.

Se tali informazioni sono sconosciute prima di monitorare il processo è consigliato utilizzare un valore moderato, ad esempio μ_{min}= 1.5, poiché ha prestazioni relativamente buone per diverse grandezze di spostamenti medi. Di conseguenza, μ_{min} è selezionato per essere 1.5 nelle applicazioni riportate in seguito.

3.2 Il vettore aumentato

Le righe di codice sottostanti hanno lo scopo di calcolare il vettore aumentato come descritto nella sezione 2.4.1.

```
augmented_vector <- function(rsada, x) {
    mu <- rsada$mu
    F <- rsada$F
    f <- rsada$f
    A <- sum(f(x-mu)/f(x), na.rm=TRUE)
    i <- which.max(x)
    xi <- x[i]
    B <- F(xi)
    C <- F(xi-mu)
    not.obs <- which(is.na(x))
    d <- length(not.obs)
    eta <- numeric(rsada$p)
    eta[i] <- (B^d*A+B^(d-1)*C*d)/(A+d)
    eta[not.obs] <- (1-eta[i])/d
    eta
}
```

Listing 3.2: vettore aumentato

Gli argomenti richiesti dalla funzione sono:

- $rsada$, una lista ritornata dalla funzione newRSADA che contiene al suo interno tutti gli elementi necessari al calcolo del vettore aumentato η;
- X, un vettore di dimensione p in cui sono considerati osservati solo gli elementi non NA.

La funzione ritorna il vettore aumentato η necessario per il calcolo della statistica test.

3.2.1 Verifica sensatezza del vettore aumentato

In questo paragrafo viene proposto un grafico per verificare la sensatezza del vettore aumentato.

A fini esplicativi viene definito p=10, q=3, μ_{min}= 1.5, k=1.

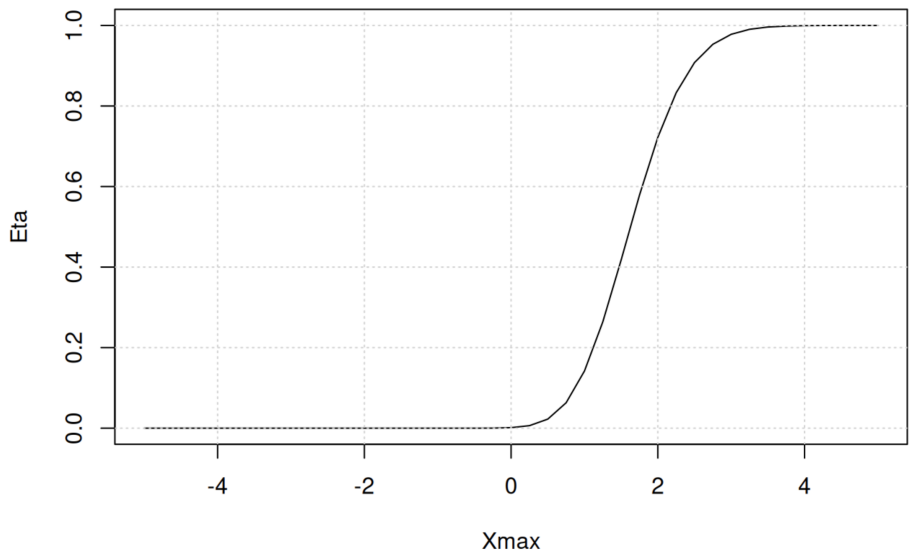

Figura 3.1: Grafico per verificare la sensatezza del vettore aumentato

Tramite il calcolo del vettore aumnetato $\eta(t)$ corrispondente al tempo t, grazie alla funzione $augmented\_vector$, viene associato a ciascuna delle p osservazioni (una per ogni variabile), realmente effettuata o potenziale, uno score.

Sull'asse delle ordinate troviamo il punteggio η che indica la probabilità che

la variabile osservata sia la più grande tra tutte le variabili (anche tra quelle non osservate). Per le variabili osservate, ma che non sono il massimo, il punteggio assegnato è logicamente 0; per quelle non osservate in egual misura viene ripartito il punteggio $(1-\eta)/(p-q)$. Sull'asse delle ascisse troviamo $Xmax$ indica il valore del massimo tra le 3 osservazioni rilevate.

Si ricorda che in base all'entità di $Xmax$ viene assegnata una probabilità coerente alla distribuzione dalla quale vengono generate le osservazioni (in questo caso normale standard); più ci si allontana dallo 0 verso valori grandi (si tratta di una carta di controllo unilaterale) più lo score associato a quell' osservazione sarà alto perchè sarà meno probabile che tra le variabili non osservate ve ne sia una che assuma un valore superiore.

Ad esempio se, come in questo caso, vengono osservate 3 variabili e il valore massimo tra queste è -2 è altamente improbabile che questo sia il massimo. Il massimo si troverà tra le altre 7 variabili e, man mano il valore massimo osservato aumenta, si avrà maggior probabilità che la variabile ad esso associata sia effettivamente la più grande e di conseguenza quella che necessita di continuare ad essere monitorata.

In conclusione il grafico proposto mostra che il vettore aumentato è sensato perchè associa a valori più elevati una maggiore probabilità di essere i valori più grandi tra tutti quelli rilevati e non; il metodo utilizzato risulta pertanto utile per sopperire all' impossibilità di rilevare tutte le osservazioni di tutte le variabili ad ogni istante t.

3.3 Aggiornamento di R-SADA

In questo paragrafo viene definita la funzione che permette il calcolo della statistica di monitoraggio discussa nel paragrafo 2.4.3.

```
updateRSADA <- function(rsada, x) {
    p <- rsada$p
    q <- sum(!is.na(x))
    g <- rsada$g
    k <- rsada$k
    eta <- augmented_vector(rsada, x)
    S1 <- rsada$S1+eta
    S2 <- rsada$S2+g
    C <- sum((S1-S2)^2/S2)
    if (C <= k) {
        S1 <- S2 <- g
```

```
12      } else {
13          C <- (C-k)/C
14          S1 <- C*S1
15          S2 <- C*S2
16      }
17      rsada$eta <- eta
18      rsada$S1 <- S1
19      rsada$S2 <- S2
20      rsada$y <- sum((S1-S2)^2/S2)
21      rsada$O <- order(-S1)[1:q]
22      rsada
23  }
```

Listing 3.3: Aggiornamento di RSADA

La funzione, proprio come quella per il calcolo del vettore aumentato, richiede in input una lista composta dagli elementi ritornati dalla funzione $newRsada$ e X, un vettore di dimensione p in cui sono considerati osservati solo gli elementi non NA.

$updateRSADA$ ha lo scopo di completare la lista degli elementi che erano solamente stati pre-allocati e posti pari a 0 con la chiamata alla funzione $newRSADA$. Le quantità interessanti che vengono calcolate sono:

- $S1$ e $S2$, le statistiche CUSUM per $\eta(t)$ e g;
- y, la statistica di monitoraggio;
- O, vettore che raccoglie le variabili che devono essere osservate al tempo t perchè considerate più sospette.

Riprendendo l'ultimo punto la funzione definisce anche in forma indiretta (grazie al calcolo di $S1$) la strategia di campionamento (si ricorda che le variabili che hanno un corrispondente valore di $S1$ elevato saranno quelle che continueranno ad essere monitorate oppure che cominceranno ad esserlo). Questo aspetto viene definito nella $riga21$ del codice dove la funzione $order$ riordina, in questo caso, gli elementi del vettore $S1$ dal più grande al più piccolo e poi vengono estratti i primi q che saranno corrispondenti alle variabili che dovranno essere osservate al tempo successivo.

In questo passaggio risulta importante notare che in presenza di ambiguità, come per esempio potrebbe essere il caso in cui si abbiano dei valori di $S1$ coincidenti, la permutazione mantiene l'ordine dell'indice dal più basso al più alto.

Capitolo 4

Applicazione del metodo R-SADA a dati simulati

In questo capitolo verrà proposta un'applicazione della procedura sinora presentata a dei dati simulati, l'esposizione e commento della carta di controllo che ne deriva e il calcolo dei limiti di controllo tramite due metodi (limiti dinamici e limiti calcolati tramite metodo bootstrap).

4.1 Inizializzazione della carta di controllo

L'esempio proposto si colloca nell'ambito di dataset complessi ma con numerosità non troppo elevata. E' stato scelto un numero di variabili potenzialmente osservabili p pari a 100 mentre quelle che saranno effettivamente osservate q per ogni istante t sono 25. Il parametro μ_{min} viene definito pari a 1.5 per le motivazioni discusse nel capitolo precedente e k è pari a 3.

$Nmax$ indica il numero massimo di iterazioni del processo di generazione dei dati e di conseguenza di campionamento.

```
p <- 100
q <- 25
mu_min <- 1.5
k <- 3
chart <- newRSADA(p, q, mu_min, k)
Nmax <- 200
```

Listing 4.1: La carta di controllo che viene utilizzata

4.2 Simulazione dei dati e calcolo della statistica di controllo

```
set.seed(12345)
tau <- 101
mu_ic <- rep(0, p)
mu_oc <- c(rep(0, 95), rep(2,5))
X <- matrix(NA, Nmax, p)
eta <- S1 <- S2 <- X
y <- numeric(Nmax)
for (i in 1:Nmax) {
    X[i, ] <- rnorm(p, if (i<tau) mu_ic else mu_oc)
    X[i, -chart$O] <- NA
    chart <- updateRSADA(chart, X[i,])
    eta[i,] <- chart$eta
    S1[i,] <- chart$S1
    S2[i,] <- chart$S2
    y[i] <- chart$y
}
```

Listing 4.2: Simulazione dati e calcolo statistica

Per la simulazione dei dati è stato definito un seed pari a 12345 in modo da rendere ripetibili i risultati ma la scelta di questo è del tutto casuale.

tau indica l'istante in cui il processo va fuori controllo, qui a t=101.

I vettori p-variati che contengono i dati vengono generati da una normale con media pari a 0 se il processo è in controllo ($t < \tau$=101), in caso contrario ($t > \tau$=101) il vettore conterrà le ultime 5 osservazioni che sono generate da una distribuzione normale con media fuori controllo ($\mu_{OC} = 2$). Il fatto che siano le ultime ad essere definite OC è stata una semplice scelta, diversamente non cambierebbero i risultati.

Una volta generato il vettore composto da 100 osservazioni vengono definite come NA quelle che non devono essere rilevate al tempo corrente, questa informazione la si ricava dal campo O.

Infine viene chiamata la funzione $updateRSADA$ che calcola il vettore aumentato, $S1$, $S2$ e la statistica di controllo y.

Viene riportata in figura 4.1 la carta CUSUM della statistica di controllo

y. In ascisse si trova l'istante temporale t mentre sull'asse delle ordinate troviamo il valore della statistica di controllo y.

Si può notare come prima del tempo t=101 ci siano dei picchi locali ma i valori della statistica tornano velocemente a 0 denotando una condizione del processo in controllo. Dal tempo t=101 i valori di y invece crescono in modo repentino senza dare l'impressione di volersi arrestare.

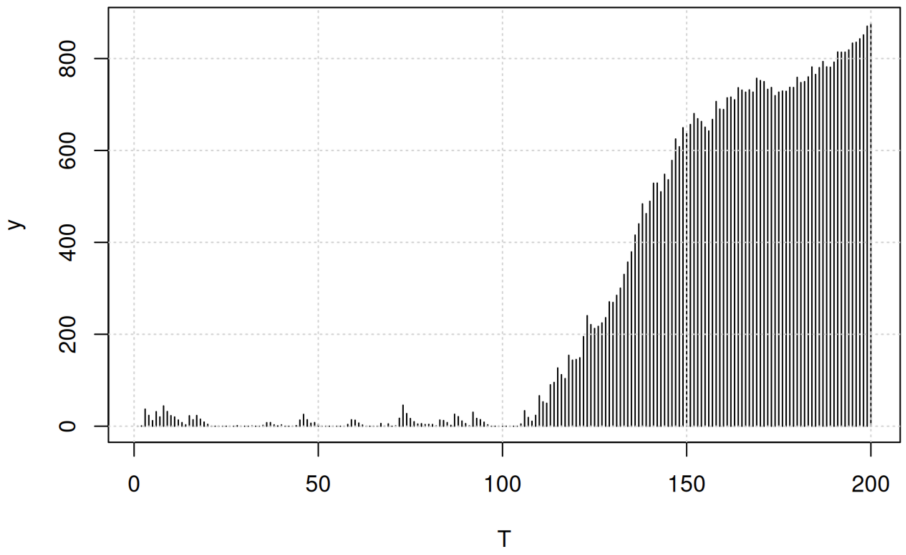

Figura 4.1: Carta CuSum

Il grafico presentato in figura 4.2 riporta sull'asse delle ascisse l'istante temporale t mentre sull'asse delle ordinate la percentuale di variabili fuori controllo che vengono campionate ad ogni istante t.

Più precisamente per ogni vettore di osservazioni si calcola la somma delle variabili campionate che sono fuori controllo o lo saranno in seguito (ovvero avranno prima o poi $\mu_{OC}>0$). Si vuole sottolineare quante delle variabili tra la 96esima e la 100esima vengono campionate per capire quanto il metodo proposto sia ottimale per rilevare le variabili da tenere sotto controllo.

Per esempio in corrispondenza di t=50 viene campionata solo una variabile tra le cinque che saranno fuori controllo e poi nessuna. Questo avviene perchè fino a che il processo è in controllo vengono osservate a rotazione tutte le variabili e siccome quelle che poi andranno fuori controllo ancora non lo sono non si continua a campionarle. Da t=101, quando il processo diventa

OC si nota che le variabili che sono da rilevare per dare l'allarme vengono campionate sempre più frequentemente con una percentuale che cresce fino a raggiungere a t=145 la totalità.

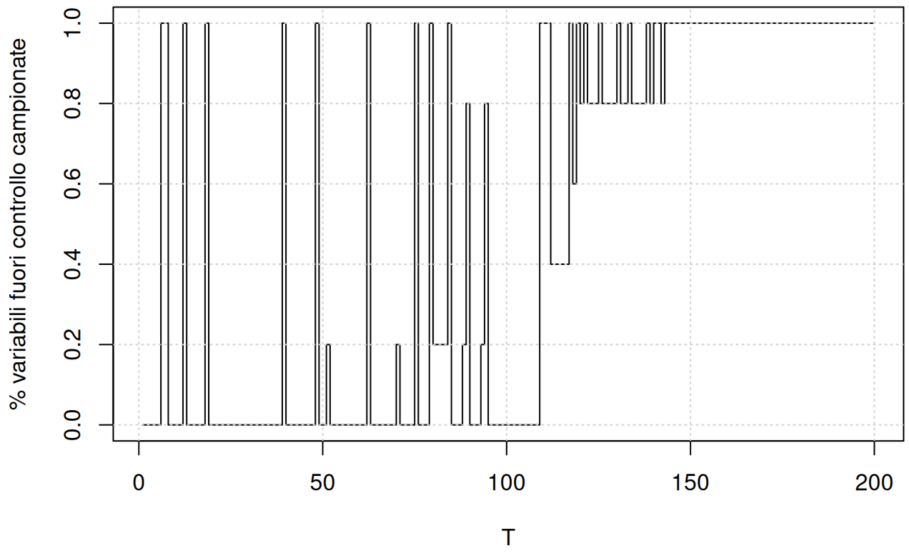

Figura 4.2: Efficacia schema di campionamento

4.3 Calcolo dei limiti

4.3.1 Limiti dinamici

Nell'articolo originale gli autori usano dei limiti di controllo costanti; viceversa, dato che la distribuzione della statistica di controllo Wt cambia con t, per la condizione di inizializzazione è stato considerato l'uso di limiti di controllo dinamici.

Utilizzare un limite di controllo variabile con t significa segnalare un allarme quando

$$Wt > Lt$$

dove Lt è una successione appropriata.

Una possibilità (valida sempre) di garantire l'ARL in controllo desiderata, consiste nel fissare Lt tale che

$$Pr(RL > t|RL >= t) = 1 - \frac{1}{B} \forall t \tag{4.1}$$

Questo infatti garantisce che la distribuzione della run length sia geometrica con media B. Dove B indica ogni quanto tempo è accettabile chiamare un falso allarme ovvero dichiarare che un processo è fuori controllo quando così non è.

Infatti,

$$Pr(RL = t) = Pr(RL > 1) \times Pr(RL > 2|RL \geq 2) \times ... \times$$

$$\times Pr(RL > t-1|RL \geq t-1) \times ...[1 - Pr(RL > t|RL \geq t)] = (1 - \frac{1}{B})^{t-1} \frac{1}{B} \tag{4.2}$$

Quindi, Lt può essere calcolato come il quantile $1 - \frac{1}{B}$ della distribuzione di Wt condizionata a $W_1 \leq L_1$, $W_2 \leq L_2$, ..., $W_{t-1} \leq L_{t-1}$, visto che

$$Pr(RL > t|RL >= t) = Pr(W_t \leq L_t|W_1 \leq L_1, \cdots, W_{t-1} \leq L_{t-1}) \tag{4.3}$$

I limiti di controllo descritti sono usualmente chiamati 'conditional false alarm control limits', o, più semplicemente limiti dinamici e sono implementabili via simulazione.

Calcolo di Lt via simulazione

La distribuzione di $W_t|W_1, ..., W_{t-1}$ è difficile da determinare analiticamente ma si possono approssimare i limiti di controllo via simulazione. Bisogna generare in parallelo alla vera statistica di controllo Wt calcolata dai dati anche un gran numero di statistiche W_t^* generate dalla distribuzione in controllo di Wt.

Un possibile algoritmo è il seguente:

Si sceglie $Nsim > 0$ e si pone

$$W_{0,1}^* = \cdots = W_{0,Nsim}^* = 0$$

al tempo t.

Simulare $Nsim$ determinazioni

$$x^*_{t,1} = \cdots = x^*_{t,Nsim}$$

dalla distribuzione in controllo di x_t. Calcolare $W^*_{t,i}$, la statistica di controllo per ogni t e per ogni i, porre poi

$$L_t = quantile\_empirico(1 - \frac{1}{B}) \text{ di } W^*_{t,1} = \cdots = W^*_{t,Nsim} \tag{4.4}$$

Infine è necessario sostituire tutte le $W^*_{t,i}$ maggiori di L_t con valori ricampionati casualmente dalle $W^*_{t,i}$, minori o uguali a L_t.

L'ultimo passo serve a far 'continuare' solo le traiettorie in controllo, ovvero ad imporre la condizione

$$W_1 \leq L_1, W_2 \leq L_2, \cdots, W_{t-1} \leq L_{t-1}$$

La bontà della approsimazione migliora all'aumentare di $Nsim$.

Funzione generica per calcolare i limiti dinamici

Prima di addentrarsi in quella che sarà la funzione per calcolare i limiti dinamici bisogna introdurre delle funzioni secondarie ma nececessarie.

- *cloneRSADA*, genera una copia della carta di controllo passata in input;
- *statRSADA*, ritorna la statistica di controllo y corrispondente alla carta di controllo passata in input;
- *rRSADA*, genera, a partire dalla distribuzione normale standard, un vettore di q osservazioni in corrispondenza delle unità da cmpionare e $p - q$ 'NA' per i dati mancanti (i limiti calcolati dipendono pertanto dallo schema di campionamento usato). Modificando questa funzione è facile generare i dati a partire da qualsiasi distribuzione.

Di seguito viene riportato il codice per calcolare i limiti dinamici.

```
1 dynamic_limits <- function(chart, clone, update, stat, rchart,
    ARL0, Tmax=50, Textra=20, Nsim=20*B) {
2     charts <- replicate(Nsim, clone(chart), simplify = FALSE)
3 # genera Nsim volte la carta al tempo 0
4     Lraw <- numeric(Tmax+Textra)
```

```
5     y <- numeric(Nsim)
6     for (i in 1:(Tmax+Textra)) {
7         for (j in 1:Nsim) {
8             charts[[j]] <- update(charts[[j]], rchart(chart))
9 # genera dei dati in controllo rchart, carte in charts
10 # aggiornate con dei dati in controllo
11             y[j] <- stat(charts[[j]])
12         }
13         Lraw[i] <- quantile(y, 1-1/B)
14         idx <- which(y>Lraw[i])
15         charts[idx] <- charts[sample(which(y<=Lraw[i]), length(
    idx))]
16 # sostituzione delle carte fuori controllo con una in controllo
17 # scegliendo in modo casuale
18     }
19     list(L=rev(isoreg(rev(Lraw))$yf)[1:Tmax], Lraw=Lraw)
20 }
```

Risulta importante notare che il processo in controllo che viene qui discusso è stazionario, anche i dati lo sono e di conseguenza anche $S1$, $S2$ e y avranno una distribuzione stazionaria. Ad un certo istante pertanto i limiti raggiungono un asintoto anche se il momento preciso dipende dalla memoria della carta.

Tali limiti, calcolati via simulazione, saranno caratterizzati da una certa variabilità ma essendo stazionari non è necessario avere un $Nsim$ troppo elevato perchè una volta ottenuti i limiti di controllo grezzi basterà calcolare la curva liscia che li interpola. Nel codice questo passaggio è svolto alla $riga15$ grazie all'utilizzo della funzione $isoreg$.

La figura 4.3 sottolinea l'efficienza del metodo che raggiunge l'asintoto tra T=10 e T=20 pertanto sarebbe possibile arrestarsi prima con le simulazioni e estendere tale valore per tutti gli istanti successivi.

In conclusione questo approccio permette di avere un'accuratezza adeguata con un numero non troppo elevato di simulazioni.

Esempio con i limiti

Ai dati simulati al paragrafo 4.2 viene applicato il metodo esposto e si riporta in figura 4.4 la carta di controllo con il metodo R-SADA e limiti dinamici.

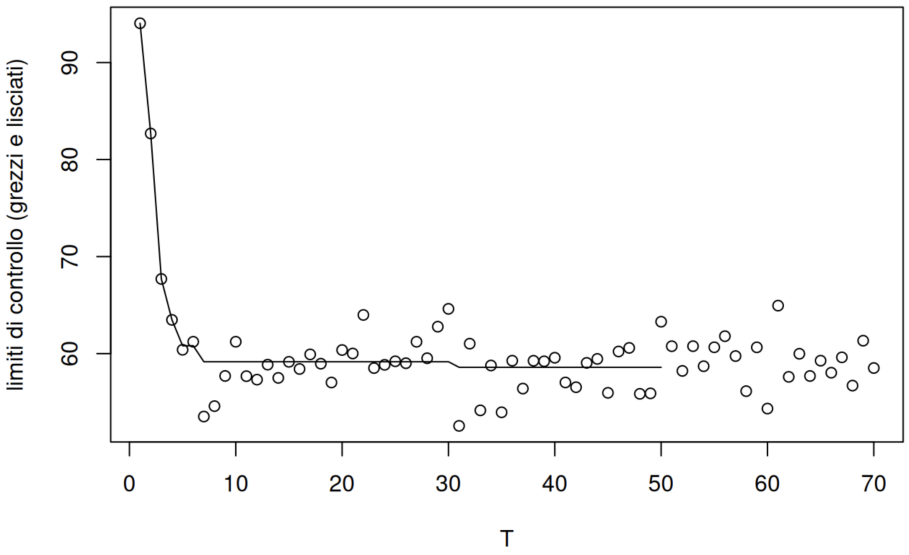

Figura 4.3: Curva liscia calcolata a partire dai limiti grezzi

Si può notare che l'allarme viene segnalato prontamente (si ricorda che il processo va fuori controllo al tempo T=101), più precisamente è stata calcolata la run length che è pari a 110 e il detection delay è pari a 10.

4.3.2 Limiti bootstrap

Per il calcolo dei limiti Lt viene proposto anche un secondo metodo tramite le stime bootstrap (stime non parametriche) mantenendo lo score η calcolato in precedenza con l'utilizzo della distribuzione Normale.

Il bootstrap è una tecnica statistica di ricampionamento con reimmissione per approssimare la distribuzione campionaria di una statistica.

In seguito viene presentata la funzione *dynamic_limits* che implementa questo metodo in cui:

- Nic, sono le osservazioni in controllo disponibili e, dato che tra i vettori di dati raccolti ai diversi istanti di tempo non vi sono differenze, vengono campionate (con reintroduzione) da questi dati senza distinguere le variabili. Nic viene definito pari a 10000 poichè p=100 è un campione di dimensione grande ma non esagerato;

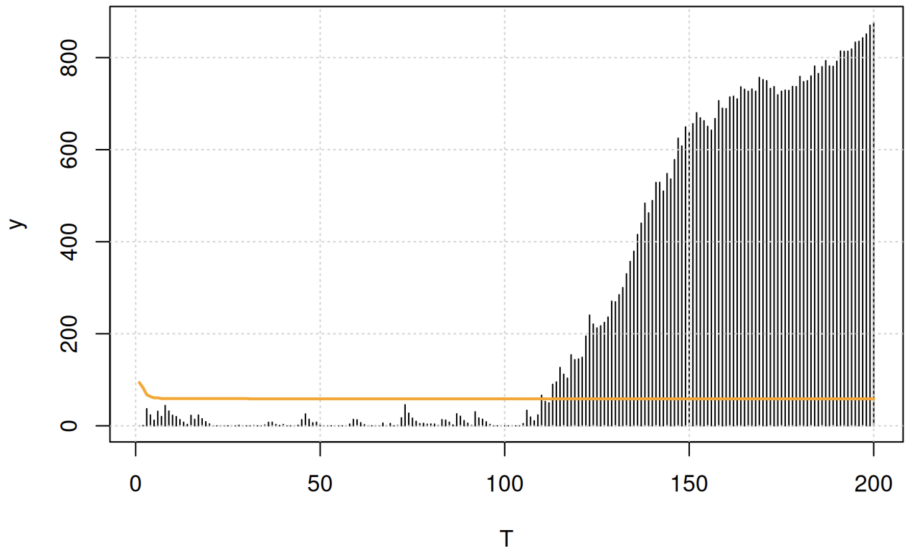

Figura 4.4: Carta di controllo R-SADA e limiti dinamici

- *Nrep*, numero di volte in cui vengono calcolati i limiti per vedere l'effetto dei dati in controllo.

```
Nic <- 10000
Nrep <- 5
Lrb <- replicate(Nrep, {Xic <- rnorm(Nic);
dynamic_limits(chart, cloneRSADA, updateRSADA, statRSADA,
     function(u) {
          x <- rep(NA, u$p)
          x[u$0] <- sample(Xic, u$q, replace=TRUE)
          x}, 500)$L})
```

L'obiettivo, oltre a proporre un metodo alternativo, era quello di capire se il metodo bootstrap genera limiti con eccessiva variabilità risultando essere una scelta poco conveniente.

Come viene mostrato nella figura 4.5 ripetendo il procedimento 5 volte c'è della variabilità ma calcolando la RL ci si accorge che in tutti i casi l'allarme viene chiamato allo stesso istante t=110 pertanto si conclude che anche l'utilizzo questo metodo potrebbe essere interessante.

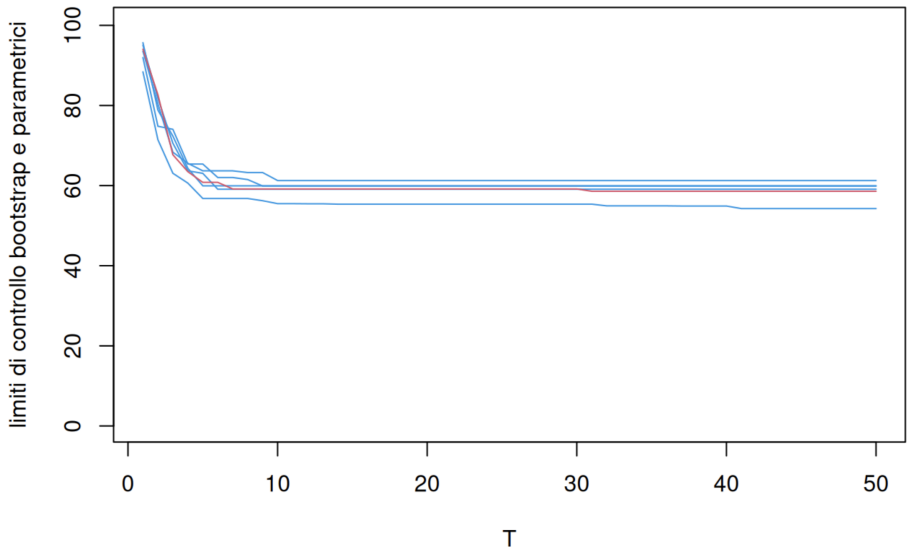

Figura 4.5: limiti di controllo bootstrap e parametrici

4.4 Un esempio bilaterale

Nell'ultima sezione del capitolo si vuole proporre per completezza un esempio in cui si vuole tenere sotto controllo le osservazioni utilizzando uno schema bilaterale. Per farlo utilizzando R-SADA è sufficiente sorvegliare i valori assoluti delle variabili standardizzate originali. Sotto un'assunzione di normalità è necessario modificare le funzioni dalle quali vengono generate le osservazioni come riportato in seguito.

```
F <- function(x) ifelse(x<=0, 0, 2*pnorm(x)-1)
f <- function(x) ifelse(x<=0, 0, 2*dnorm(x))
```

Sensatezza del vettore aumentato

In figura 4.6 viene verificata la sensatezza dello score η in uno schema bilaterale. Il principio risulta essere sempre lo stesso: allontanandosi dallo 0 in entrambe le direzioni la probabilità che l'osservazione $Xmax = \left\{ x_i | \ |x_i| = \underset{j=1,\dots,q}{max} |x_j^O| \right\}$ sia la più grande tra tutte cresce in modo proporzionale rendendola sempre più sospetta.

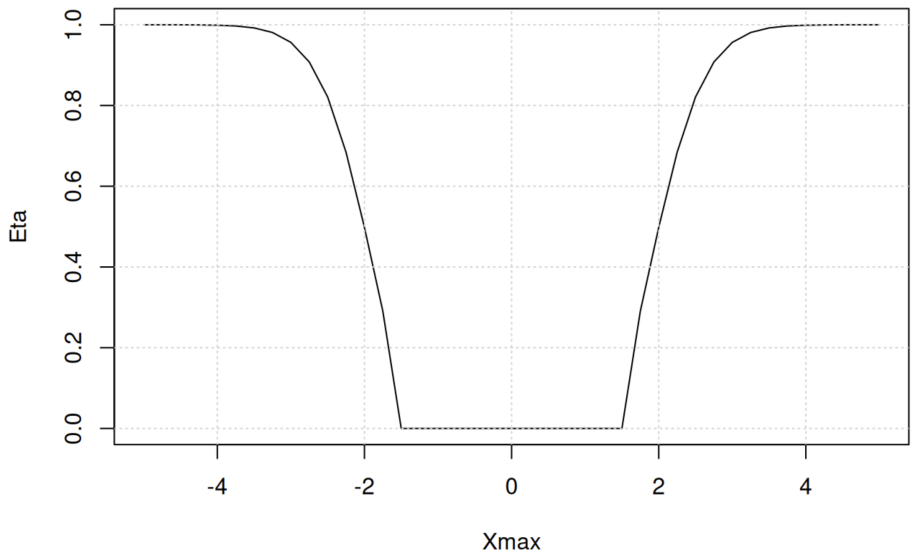

Figura 4.6: Grafico per verificare la sensatezza del vettore aumentato

Calcolo della carta di controllo

Il codice seguente ha lo scopo di calcolare la statistica di controllo e definire lo schema di campionamento. Come per il caso unilaterale è stato scelto un numero di variabili potenzialmente osservabili pari a 100 (p) mentre quelle che saranno effettivamente osservate per ogni istante t sono 25 (q). Il parametro μ_{min} viene definito pari a 1.5 e k è pari a 3.

```
p <- 100
q <- 25
mu_min <- 1.5
k <- 3
chart <- newRSADA(p, q, mu_min, k, F, f)
Nmax <- 200
set.seed(54321)
tau <- 101
mu_ic <- rep(0, p)
mu_oc <- c(rep(-3,3), rep(0, 94), rep(3, 3))
X <- matrix(NA, Nmax, p)
```

```
eta <- S1 <- S2 <- X
y <- numeric(Nmax)
O <- matrix(NA, Nmax, q)
for (i in 1:Nmax) {
    X[i, ] <- rnorm(p, if (i<tau) mu_ic else mu_oc)
    X[i, -chart$O] <- NA
    chart <- updateRSADA(chart, abs(X[i,]) )
    eta[i,] <- chart$eta
    S1[i,] <- chart$S1
    S2[i,] <- chart$S2
    y[i] <- chart$y
}
```

La carta di controllo viene mostrata in figura 4.7. Si nota come dall'istante T=101 i valori di y cominciano a crescere in modo repentino senza mai tornare verso lo zero, indice che il processo è fuori controllo.

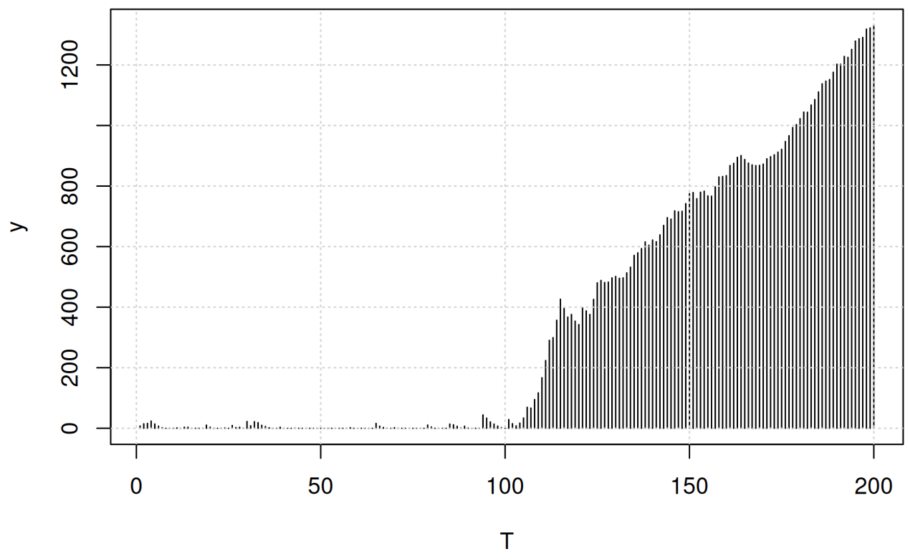

Figura 4.7: Carta di Controllo bilaterale

Nella figura 4.8 invece viene mostrato lo schema di campionamento che

anche in questo caso risulta efficace perchè non appena il processo comincia ad essere fuori controllo le variabili che hanno una distribuzione diversa da quella in controllo vengono rilevate e continuano ad esserlo fino alla fine del tempo di osservazione.

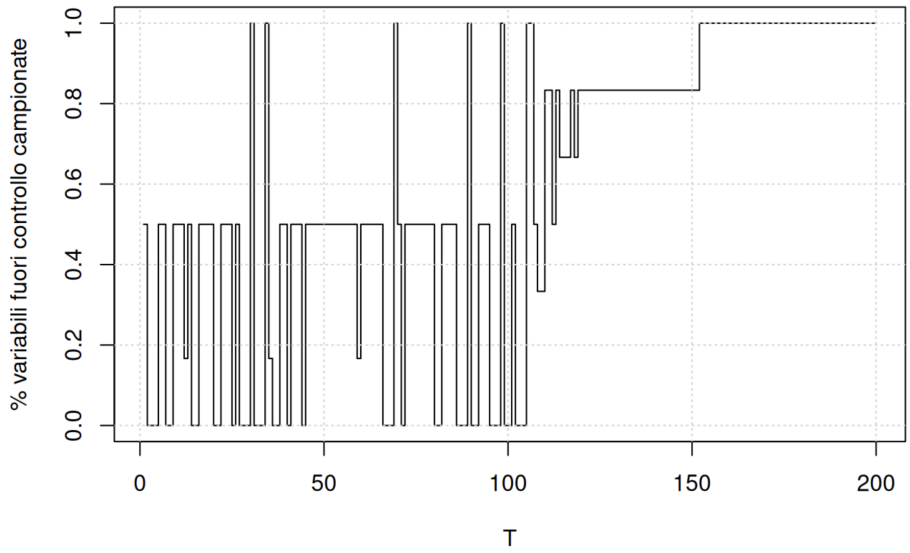

Figura 4.8: Efficacia schema di campionamento

Limiti dinamici

Per il calcolo dei limiti dinamici è necessario apportare delle modifiche ad una delle funzioni ausiliarie: $rSADA$. Si ricorda che $rRSADA$, genera, a partire dal valore assoluto della distribuzione normale standard, un vettore di q osservazioni in corrispondenza delle unità da campionare, $p-q$ 'NA' per i dati mancanti e ritorna il vettore di osservazioni x.

Di seguito il codice.

```
rSADA <- function(u) {
    x <- rep(NA, rsada$p)
    x[u$0] <- abs(rnorm(u$q))
    x
}
```

La chiamata alla funzione risulta in questo modo identica al caso unilaterale come lo è anche il grafico riportato in figura 4.9.

```
Lr <- dynamic_limits(chart, cloneRSADA, updateRSADA, statRSADA,
    rRSADA, 500)
```

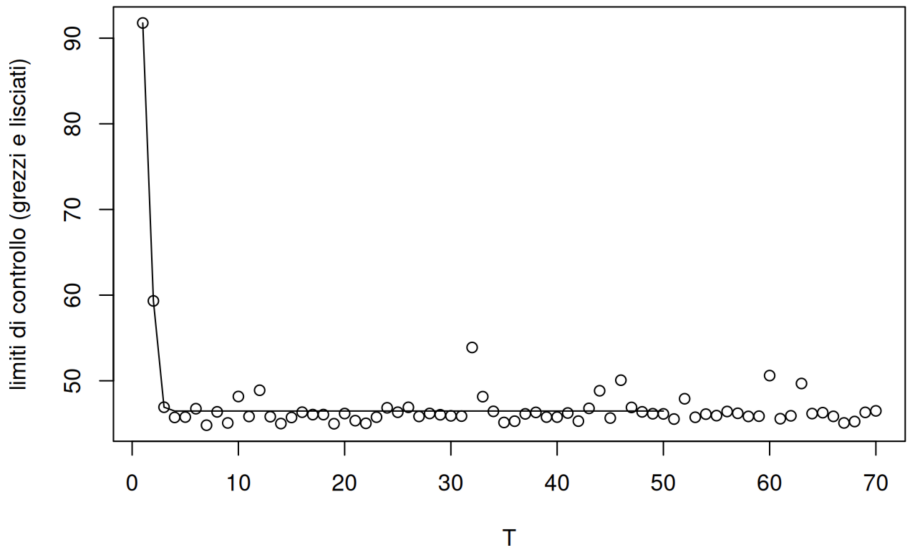

Figura 4.9: limiti di controllo (grezzi e lisciati)

Per concludere in figura 4.10 si trova la carta di controllo calcolata con il metodo R-SADA e limiti dinamici.

Si nota che l'allarme viene segnalato prontamente, la RL è pari a 106 e il DD è pari a 6.

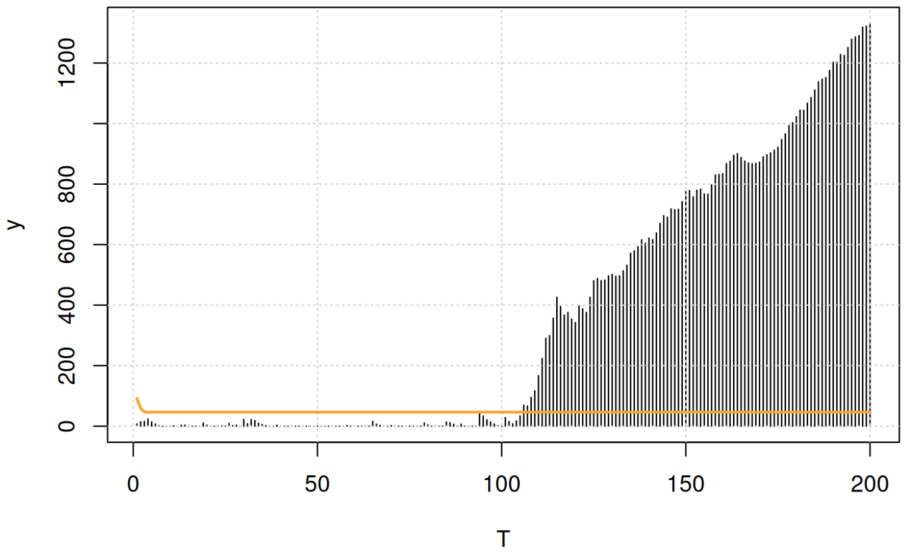

Figura 4.10: Efficacia schema di campionamento

Capitolo 5

Uno schema di campionamento alternativo

In questo capitolo verrà presentato un problema riscontrato durante l'applicazione della carta di controllo R-SADA legato allo schema di campionamento.

5.1 I limiti della Carta di Controllo proposta

Nel paragrafo 2.4.4 viene presentata la "IC property": tutte le variabili sono visitate prima o poi. L'algoritmo però campiona solo le variabili più sospette, quelle che una volta calcolata la statistica di controllo sono tra quelle che presentano l' $S1$ più elevato.

Se viene ridotta in modo significativo la frazione campionata può capitare che il metodo reagisca con molto ritardo alla situazione fuori controllo. Questo avviene perchè le vere variabili fuori controllo non sono campionate per un lasso di tempo rilevante, ovvero, è possibile che l'algoritmo rimanga incastrato nell'esplorare variabili che sospetta siano fuori controllo ma che in realtà non lo sono.

Quando lo schema di controllo viene applicato ad un insieme di dati può avvenire, ed è molto frequente, che variabili diverse presentino lo stesso valore di $S1$ e quando poi vengano definite quelle da rilevare vengono preferite quelle che in partenza occupavano le prime posizioni non trattandosi così di un'estrazione casuale.

5.1.1 Esempio

Per mostrare il problema si utilizzano gli stessi dati unilaterali del Capitolo 4 con l'unica differenza che q viene definito pari a 10. Di seguito viene riportato il codice per il calcolo della statistica di controllo e la definizione dello schema di campionamento.

```
p <- 100
q <- 10
chart <- newRSADA(p, q, mu_min, k)
Nmax <- 400
set.seed(12345)
tau <- 101
mu_ic <- rep(0, p)
mu_oc <- c(rep(0, 95), rep(2,5))
X <- matrix(NA, Nmax, p)
eta <- S1 <- S2 <- X
y <- numeric(Nmax)
for (i in 1:Nmax) {
    X[i, ] <- rnorm(p, if (i<tau) mu_ic else mu_oc)
    X[i, -chart$O] <- NA
    chart <- updateRSADA(chart, X[i,])
    eta[i,] <- chart$eta
    S1[i,] <- chart$S1
    S2[i,] <- chart$S2
    y[i] <- chart$y
}
```

Come si può notare dalla figura 5.1 la carta di controllo non reagisce in modo repentino quando il processo va fuori cotrollo in corrispondenza di T=101 ma comincia a crescere con molto ritardo.

In figura 5.2 invece risulta chiaro il limite del metodo proposto. Fino all'istante T=355 nessuna variabile tra quelle responsabili del malfunzionamento del processo viene osservata e questo influisce pesantemente sull'istante in cui viene chiamato l'allarme.

I limiti sono stati calcolati con l'utilizzo dei limiti dinamici e vengono riportati insieme alla carta di controllo completa in figura 5.3.

A riprova di ciò che è stato visualizzato nei grafici è stata calcolata la RL che risulta pari a 356 mentre il DD è pari a 256.

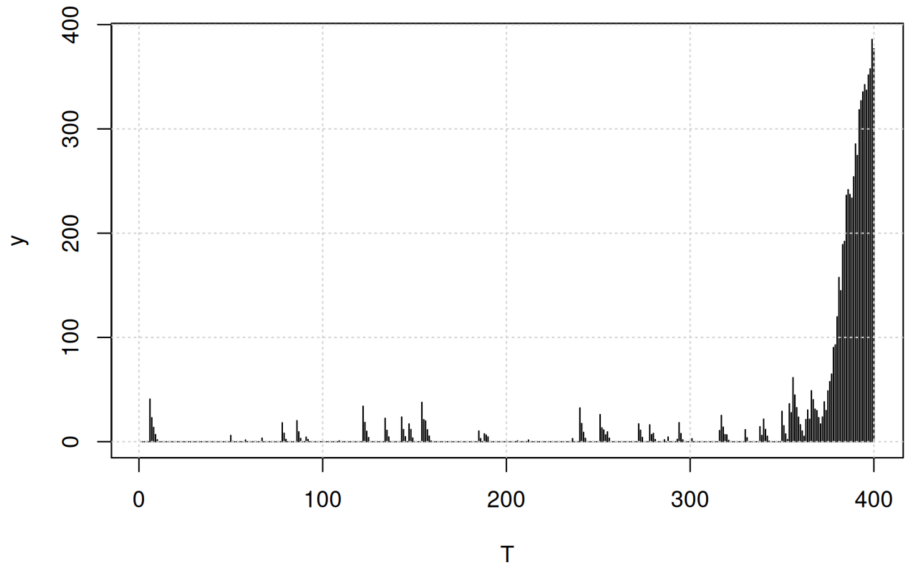

Figura 5.1: Carta di Controllo R-SADA

5.2 Uno schema di campionamento alternativo

Una procedura di campionamento alternativa che risolve il problema precedente consiste nel campionare:

- le $q1$ variabili più sospette ovvero quelle che sono caratterizzate da S1 grande;
- le $q2$ variabili che non sono state visitate da più tempo.

Operando in questo modo se se $p/q2$ non è particolarmente elevato ogni variabile non è mai tralasciata per un lungo periodo.

Si noti che se $q1+q2 = q$, dove q indica il numero di variabili visitate dallo schema originale, il costo di campionamento di questo schema è lo stesso di quello precedente.

5.2.1 Implementazione

In seguito viene presentata una possibile implementazione dello schema di campionamento sopra proposto.

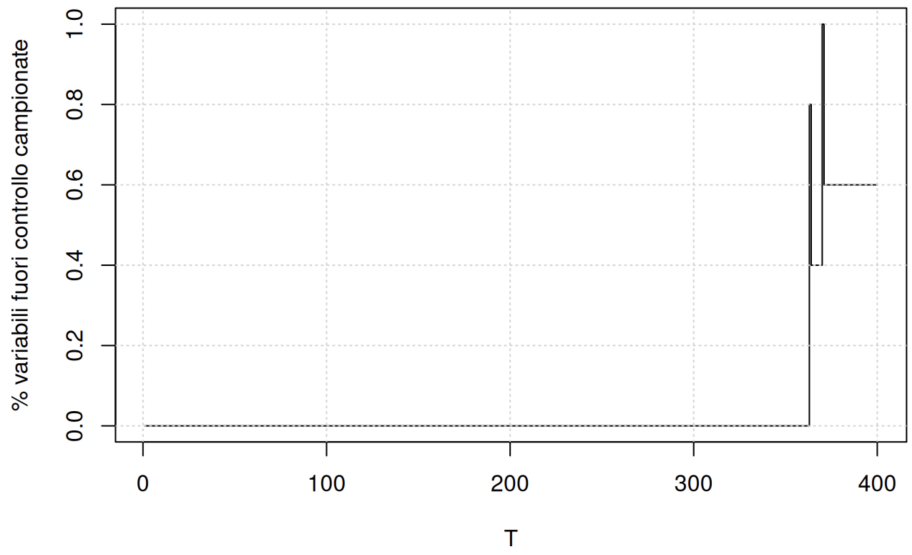

Figura 5.2: Efficacia schema di campionamento

Per semplicità e per rendere visibili le differenze rispetto al precedente modo di operare vengono riutilizzate le funzioni già discusse con alcune modifiche.

```
1 newRSADA1 <- function(p, q1, q2, mu_min, k, F=pnorm, f=dnorm)
    {
2     a <- newRSADA(p, q1+q2, mu_min, k, F, f)
3     a$q1 <- q1
4     a$q2 <- q2
5     a$l <- integer(p)
6     a
7 }
```

*newRSADA*1 è la funzione che permette di inizializzare la carta di controllo e corrisponde a questa all'istante 0. Riceve come argomenti gli stessi di *newRSADA* introdotti nel paragrafo 3.1 con la differenza che, al posto del campo q, vengono passati $q1$, $q2$ e viene definito il campo l in cui viene memorizzato da quanto tempo le variabili non vengono visitate.

```
1 updateRSADA1 <- function(rsada, x) {
2     rsada <- updateRSADA(rsada, x)
```

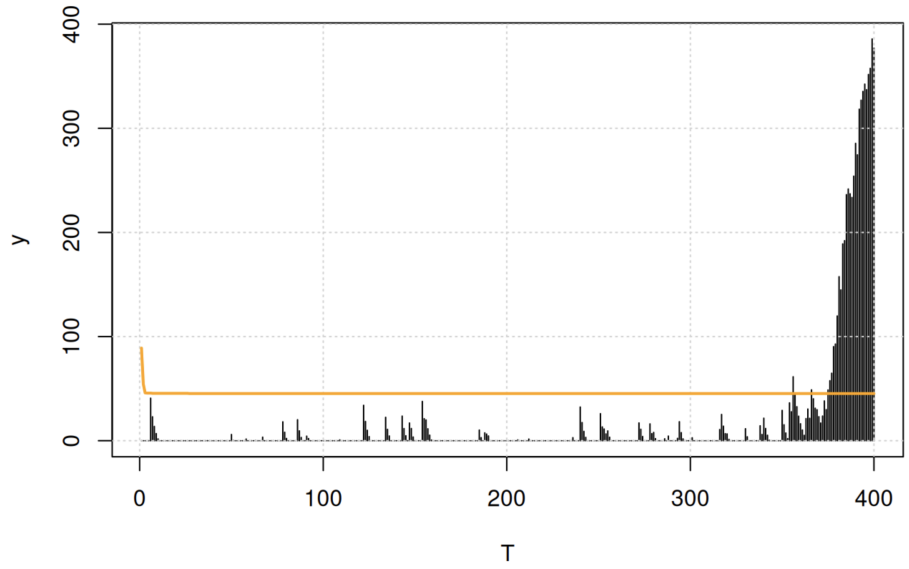

Figura 5.3: Efficacia schema di campionamento

```
3     rsada$l <- rsada$l + 1
4     rsada$l[!is.na(x)] <- 0
5     rsada$O <- c(order(-rsada$S1)[1:rsada$q1],
6                  order(-rsada$l)[1:rsada$q2])
7     rsada
8 }
9 cloneRSADA1 <- function(u) newRSADA1(u$p, u$q1, u$q2, u$mu, u$k
      , u$F, u$f)
10 rRSADA1 <- rRSADA
11 statRSADA1 <- statRSADA
```

L'aggiornamento della carta di controllo R-SADA, con al suo interno il calcolo del vettore aumentato, avviene tramite la funzione $updateRSADA1$ in cui le differenze dalla versione iniziale sono:

- l'aggiornamento del campo l ($riga3$), vettore in cui per ogni variabile non osservata viene aggiunto un ritardo pari a 1 mentre quelle osservate vengono poste a 0 poiché appena controllate;
- il campo O ($riga5$) che definisce lo schema di campionamento viene aggiornato inserendovi le $q1$ variabili più sospette (con $S1$ più elevato) e

allo stesso tempo vengono aggiunte anche le $q2$ variabili meno osservate (al più una variabile non la posso osservare per $p/q2$ istanti di tempo).

Le funzioni $cloneRSADA1$, $rSADA1$, $statRSADA1$ restano analoghe.

5.3 Applicazione

In questo paragrafo viene mostrato un esempio di applicazione di quanto sopra introdotto. Nel codice riportato di seguito vengono definiti p pari a 100, $q1$ pari a 5, $q2$ pari a 5 (quindi q pari a 10), μ_{min} pari a 1.5, k pari a 3.

Il processo va fuori controllo all'istante $\tau = 101$.

```
chart1 <- newRSADA1(p, 5, 5, mu_min, k)
set.seed(12345)
X <- matrix(NA, Nmax, p)
eta <- S1 <- S2 <- X
y <- numeric(Nmax)
O <- matrix(NA, Nmax, q)
for (i in 1:Nmax) {
    X[i, ] <- rnorm(p, if (i<tau) mu_ic else mu_oc)
    X[i, -chart1$O] <- NA
    chart1 <- updateRSADA1(chart1, X[i,])
    eta[i,] <- chart1$eta
    S1[i,] <- chart1$S1
    S2[i,] <- chart1$S2
    y[i] <- chart1$y
}
```

Nella figura 5.4 viene riportata la carta di controllo R-SADA. Si nota che la statistica di controllo y comincia a crescere all'incirca a partire dall'istante 110, reagisce prontamente al cambiamento che manda fuori controllo il processo.

Interessante è anche la figura 5.5 in cui vengono riportate le percentuali delle variabili OC che vengono osservate ad ogni istante. Si può notare che nello schema alternativo ogni variabile viene visitata almeno ogni 20 istanti di tempo; è quindi impossibile ignorare le variabili fuori controllo per un lungo periodo. Nel grafico questo è chiaramente visibile perchè tali variabili sono periodicamente visitate anche prima di τ, quando sono ancora in controllo.

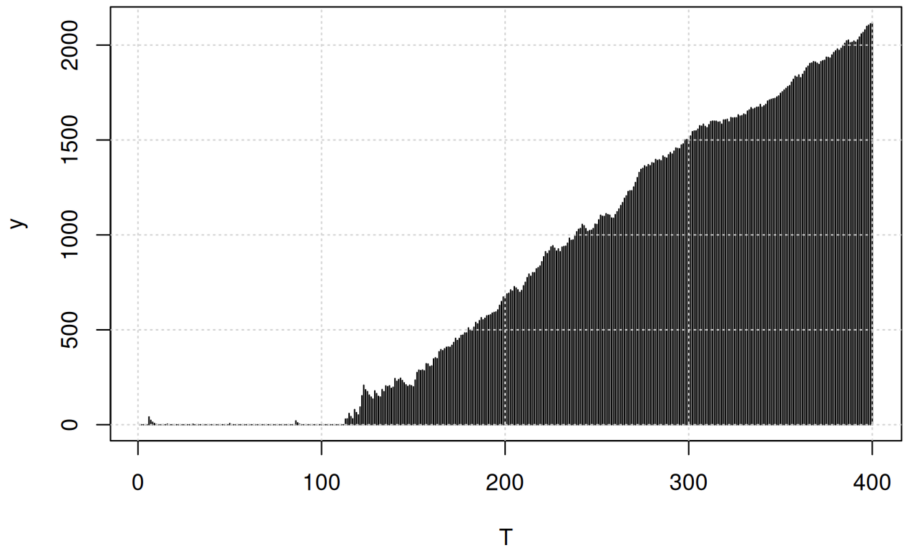

Figura 5.4: Carta di Controllo RSADA

I limiti vengono calcolati come in precedenza tramite limiti dinamici e lisciamento. La Figura 5.1 rappresenta la carta di controllo R-SADA completa di limiti dinamici. L'allarme viene prontamente chiamato quando il processo diventa fuori controllo, più precisamente si rileva una RL pari a 115 e un DD pari a 15.

```
Lr <- dynamic_limits(chart1, cloneRSADA1, updateRSADA1,
    statRSADA1, rRSADA1, 500)
```

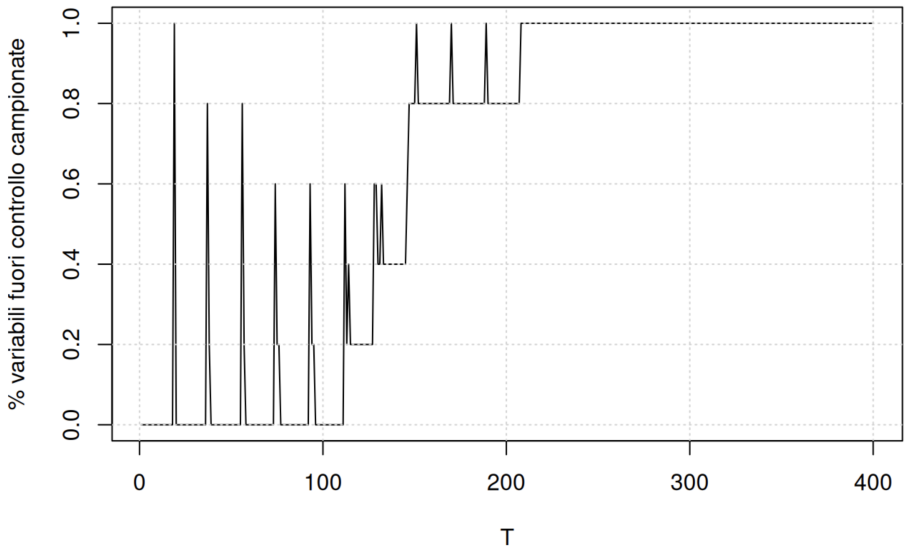

Figura 5.5: Efficacia schema di campionamento

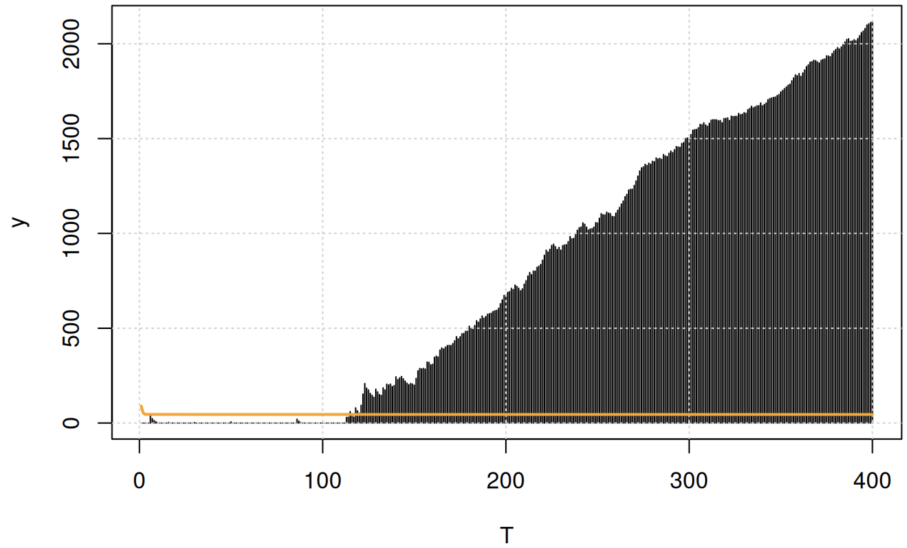

Figura 5.6: Carta di Controllo R-SADA con limiti dinamici

Conclusioni

Negli ultimi anni i vincoli di risorse sono stati ben riconosciuti come una sfida essenziale durante il monitoraggio online dei flussi di big data. Senza l'accesso alle informazioni complete sul processo è di fondamentale importanza allocare efficacemente le risorse di monitoraggio poiché gli scenari OC possono essere molto complicati e difficili da notare.

In questa relazione è stato studiato un algoritmo di campionamento basato su ranghi e sull'aumento dei dati per rilevare rapidamente gli spostamenti medi in un processo quando è disponibile solo una porzione limitata di osservazioni per ogni momento di acquisizione. L'idea è quella di facilitare l'allocazione delle risorse alle variabili OC così da ottenere un rapido rilevamento del cambiamento del processo e segnalare un'allarme.

Per dimostrare il vantaggio del metodo proposto è stato poi condotto uno studio di simulazione e ai dati è stata applicata la carta di controllo studiata. Il metodo risulta vantaggioso ma osservandolo più da vicino presenta dei limiti soprattutto nel caso in cui venga ridotta in modo significativo la frazione di variabili campionata. E' stato perciò proposto uno schema di campionamento alternativo che permette di non lasciare per troppo tempo una variabile non osservata.

Ci sono alcuni argomenti da approfondire in relazione al metodo R-SADA proposto. Ad esempio la procedura tramite aumento dei dati non risulta generale perchè si trova il vincolo che definisce le variabili come indipendenti e identicamente distribuite. Questa assunzione risulta molto forte e poco realistica.

La carta di controllo proposta è anche meno sensibile a piccoli spostamenti medi; bisognerebbe esplorare una strategia migliore per aumentare le variabili non osservabili per il rilevamento di questi.

Bisogna infine sottolineare il fatto che il metodo appare in una certa misura poco verosimile anche perchè nel calcolo del vettore aumentato si assume la conoscenza della distribuzione che ha generato le osservazioni; questo risulta molto raro quando si ha a che fare con un grande numero di variabili. Per cercare di superare la dipendenza da distribuzione parametrica

che si trova nell'articolo è stato proposto il calcolo dei limiti con metodo bootstrap che si dimostra essere un'alternativa valida purché il numero di osservazioni disponibile in fase 1 sia sufficiente.

Nonostante questo nel contesto dei big data risulterebbe più utile avere un metodo che includa la stima non parametrica delle distribuzioni che hanno generato le osservazioni tramite l'utilizzo dei dati a disposizione.

Bibliografia

[1] Adelchi Azzalini e Bruno Scarpa. *Data analysis and data mining: An introduction*. OUP USA, 2012.

[2] Ana Marıa Estrada Gómez, Dan Li e Kamran Paynabar. «An Adaptive Sampling Strategy for Online Monitoring and Diagnosis of High-Dimensional Streaming Data». In: *Technometrics* 64.2 (2022), pp. 253–269.

[3] Douglas C Montgomery. *Introduction to statistical quality control*. John Wiley Sons, 2020.

[4] Mohammad Nabhan, Yajun Mei e Jianjun Shi. «Correlation-based dynamic sampling for online high dimensional process monitoring». In: *Journal of Quality Technology* 53.3 (2021), pp. 289–308.

[5] Andi Wang, Xiaochen Xian, Fugee Tsung e Kaibo Liu. «A spatial-adaptive sampling procedure for online monitoring of big data streams». In: *Journal of Quality Technology* 50.4 (2018), pp. 329–343.

[6] William H Woodall e Matoteng M Ncube. «Multivariate CUSUM quality-control procedures». In: *Technometrics* 27.3 (1985), pp. 285–292.

[7] Xiaochen Xian, Andi Wang e Kaibo Liu. «A nonparametric adaptive sampling strategy for online monitoring of big data streams». In: *Technometrics* 60.1 (2018), pp. 14–25.

[8] Xiaochen Xian, Chen Zhang, Scott Bonk e Kaibo Liu. «Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation». In: *Journal of Quality Technology* 53.2 (2021), pp. 135–153.