Tracking Predictable Drifting Parameters of a Time Series

Joint work with Eduard Belitser

Paulo Serra

Department of Mathematics and Computer Science
Eindhoven University of Technology

12th September, 2013

Model Description

Assume we observed $\boldsymbol{X}_n = (X_0, X_1, \ldots, X_n)$ following:

$$X_0 \sim \mathbb{P}_0, \qquad X_k | \boldsymbol{X}_{k-1} \sim \mathbb{P}_k(\cdot|\boldsymbol{X}_{k-1}), \quad k \in \mathbb{N}.$$

- X_k takes values in $\mathcal{X} \subseteq \mathbb{R}^l$, $l \in \mathbb{N}$ (i.e. $\mathbb{P}(X_k \in \mathcal{X}) = 1$).
- The distribution of $\boldsymbol{X}_n$, $n \in \mathbb{N}_0$, is given by
$$\mathbb{P}^{(n)} = \mathbb{P}^{(n)}(\boldsymbol{x}_n) = \prod_{k=0}^{n} \mathbb{P}_k(x_k|\boldsymbol{x}_{k-1}), \boldsymbol{x}_k \in \mathcal{X}^{k+1},$$
where $\mathbb{P}_0(x_0|\boldsymbol{x}_{-1})$ should be understood as $\mathbb{P}_0(x_0)$.
- At time $n \in \mathbb{N}_0$, the underlying growing statistical model is $\mathcal{P}^{(n)} = \left\{ \prod_{k=0}^{n} \mathbb{P}_k(x_k|\boldsymbol{x}_{k-1}) : \mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1}) \in \mathcal{P}_k \right\}$.

Objective

- Consider a filtration $\{\mathcal{F}_k\}_{k=-1}^{\infty}$ such that $\mathcal{F}_k \subseteq \sigma(\boldsymbol{X}_k)$.
- Consider a sequence of appropriately measurable operators A_k, which map measures $\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1}) \in \mathcal{P}_k$

$$A_k(\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})) = \theta_k(\boldsymbol{x}_{k-1}), \quad \boldsymbol{x}_{k-1} \in \mathcal{X}^{k-1},$$

with $\theta_k(\boldsymbol{x}_{k-1})$ predictable with respect to $\mathcal{F}_k$.

Objective

We would like to track $\theta_k = \theta_k(\boldsymbol{X}_{k-1})$.

Definition of the Algorithm

Assumption (A0)

The drifting parameter satisfies $\mathbb{P}(\theta_k(\boldsymbol{X}_{k-1}) \in \Theta)$ for some Θ such that $\sup_{\theta\in\Theta} \|\theta\|^2 \leq C_\Theta$.

The following algorithm constitutes our tracking sequence.

Tracking Algorithm

Define $\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k G_k(\hat{\theta}_k, \boldsymbol{X}_k)$, $k \in \mathbb{N}_0$ where $0 \leq \gamma_k \leq \Gamma$ and arbitrary $\mathcal{F}_{-1}$-measurable $\hat{\theta}_0 \in \Theta \subset \mathbb{R}^d$.

The functions $G_k(\hat{\theta}_k, \boldsymbol{X}_k)$ are called gain vectors.

Assumptions on the Gain Function

Assumption (A1)

For all $k \in \mathbb{N}_0$, constants λ_1, λ_2 and $\theta_k = A_k\big(\mathbb{P}_k(\cdot|\boldsymbol{x}_{k-1})\big)$,

$$g_k(\hat{\theta}_k, \theta_k) = g_k(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1}) = \mathbb{E}\big[G_k(\hat{\theta}_k, \boldsymbol{X}_k)|\mathcal{F}_{k-1}\big],$$

exists; for a $\mathcal{F}_{k-1}$-measurable symmetric PD matrix $\boldsymbol{M}_k$, a.s.

$$g_k(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1}) = -\boldsymbol{M}_k(\hat{\theta}_k - \theta_k),$$

$$0 < \lambda_1 \leq \mathbb{E}[\lambda_{(1)}(\boldsymbol{M}_k)|\mathcal{F}_{k-2}] \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2 < \infty.$$

Assumption (A2)

There exists a constant $C_g > 0$ such that

$$\mathbb{E}\|G_k(\hat{\theta}_k, \boldsymbol{X}_k) - g_k(\hat{\theta}_k, \theta_k|\boldsymbol{X}_{k-1})\|^2 \leq C_g, \quad k \in \mathbb{N}_0.$$

Theorem (Bound on L_1 risk)

Let (A0) – (A2) hold and $\delta_k = \delta_k(\boldsymbol{X}_{k-1}) = \hat{\theta}_k - \theta_k$, $k \in \mathbb{N}_0$. *Then for any* $k_0, k \in \mathbb{N}_0$ *and sequence* $\{\gamma_k, k \in \mathbb{N}_0\}$ *(satisfying the conditions of the previous lemma) such that* $\gamma_i \lambda_2 \leq 1$, $i \in \{k_0, \ldots, k\}$, *the following relation holds:*

$$\mathbb{E}\|\delta_{k+1}\| \leq C_1 \exp\Big\{ -\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_i\Big\} + C_2\Big[\sum_{i=k_0}^{k}\gamma_i^2\Big]^{1/2} + \\ + C_3 \max_{k_0 \leq i \leq k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|, \quad k_0 \leq k,$$

where $C_1 = \sqrt{2}(\bar{C}_\Theta + C_\Theta)^{1/2}$, $C_2 = C_g^{1/2}(1 + \lambda_2/\lambda_1)$, $C_3 = (1 + \lambda_2/\lambda_1)$.

Stronger Assumptions on the Gain Function

Assumption (A1)

$$g_k(\hat{\theta}_k, \theta_k | \boldsymbol{X}_{k-1}) = -\boldsymbol{M}_k(\hat{\theta}_k - \theta_k).$$

$$0 < \lambda_1 \leq \mathbb{E}[\lambda_{(1)}(\boldsymbol{M}_k)|\mathcal{F}_{k-2}] \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2 < \infty, \quad \text{(a.s.)}$$

$$\downarrow$$

$$0 < \lambda_1 \leq \lambda_{(1)}(\boldsymbol{M}_k) \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2 < \infty, \quad \text{(a.s.)}$$

Assumption (A2)

$$\mathbb{E}\|G_k(\hat{\theta}_k, \boldsymbol{X}_k) - g_k(\hat{\theta}_k, \theta_k | \boldsymbol{X}_{k-1})\|^2 \leq C_g, \quad k \in \mathbb{N}_0.$$

$$\downarrow$$

$$\|G_k(\hat{\theta}_k, \boldsymbol{X}_k)\|^2 \leq C_g, \quad k \in \mathbb{N}_0, \text{ (a.s.)}$$

Theorem (L_p risk bound)

Suppose that the conditions of the previous theorem are fulfilled. If, in addition (to assumption (A1)), $\lambda_{(1)}(\boldsymbol{M}_i) \geq \lambda_1$ *and* $\|G_i(\hat{\theta}_i, \boldsymbol{X}_i)\| \leq C_g$ *(instead of (A2)) a.s. for all* $i = k_0 \ldots, k$, *then for any* $p \geq 1$

$$\mathbb{E}\|\delta_{k+1}\|_p^p \leq C_1' \exp\Big\{-p\lambda_1 \sum_{i=k_0}^{k} \gamma_i\Big\} + C_2' \Big[\sum_{i=k_0}^{k} \gamma_i^2\Big]^{p/2} + C_3' \max_{k_0 \leq i \leq k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p^p, \quad k_0 \leq k,$$

for $C_1' = 3^{p-1} K_p^p \mathbb{E}\|\delta_{k_0}\|_p^p$, $C_2' = 3^{p-1} 2^p d B_p C_g^p (1 + K_p^2 \lambda_2/\lambda_1)^p$, $C_3' = 3^{p-1}(1 + K_p^2 \lambda_2/\lambda_1)^p$, *and* K_p *and* B_p *are constants.*

Lipschitz Signal: $\theta_k^n = \vartheta(k/n), \quad \vartheta(\cdot) \in \mathcal{L}_\beta$

$$\mathbb{E}\|\delta_{k+1}\| \lesssim \exp\Big\{-\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_i\Big\} + \Big[\sum_{i=k_0}^{k}\gamma_i^2\Big]^{1/2} + \max_{k_0\le i\le k}\mathbb{E}\|\theta_{i+1}-\theta_{k_0}\|,$$

$$\mathbb{E}\|\delta_{k+1}\|_p^p \lesssim \exp\Big\{-p\lambda_1\sum_{i=k_0}^{k}\gamma_i\Big\} + \Big[\sum_{i=k_0}^{k}\gamma_i^2\Big]^{p/2} + \max_{k_0\le i\le k}\mathbb{E}\|\theta_{i+1}-\theta_{k_0}\|_p^p,$$

$$X_0^n \sim P_{\theta_0^n}, \qquad X_k^n | \boldsymbol{X}_{k-1}^n \sim \mathbb{P}_{\theta_k^n}(\cdot|\boldsymbol{X}_{k-1}^n), \quad k \le n \in \mathbb{N},$$

- Assume that $\theta_k^n = \vartheta(k/n)$, with $\vartheta(\cdot) \in \mathcal{L}_\beta$, $k = 1, \dots, n$.
- For $0 < \beta \le 1$, $\gamma_k \equiv C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)}$, $k_0 = K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$ we get

$$\sup_{\substack{\vartheta\in\mathcal{L}_\beta\\ k\ge K_n}} \mathbb{E}\frac{n^{-\frac{\beta}{2\beta+1}}\|\delta_k\|}{(\log n)^{\frac{2\beta}{2\beta+1}}} \le C \quad \text{and} \quad \sup_{\substack{\vartheta\in\mathcal{L}_\beta\\ k\ge K_n}} \mathbb{E}\bigg(\frac{n^{-\frac{\beta}{2\beta+1}}\|\delta_k\|_p}{(\log n)^{\frac{2\beta}{2\beta+1}}}\bigg)^p \le C.$$

Example 1

Signal + noise setting

The model is:

$$X_k = \theta_k + \xi_k, \quad k \in \mathbb{N}_0,$$

where $\{\theta_k\}_{k\in\mathbb{N}_0}$ is a predictable process ($\theta_k = \theta_k(\boldsymbol{X}_{k-1})$), $\{\xi_k\}_{k\in\mathbb{N}_0}$ is a martingale difference noise with respect to the filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}_{-1}}$.

We can simply take the following gain function

$$G_k(\hat{\theta}_k, \boldsymbol{X}_k) = -(\hat{\theta}_k - X_k), \quad k \in \mathbb{N}_0,$$

since

$$g_k(\hat{\theta}_k, \theta_k | \boldsymbol{X}_{k-1}) = \mathbb{E}[G_k(\hat{\theta}_k, \boldsymbol{X}_k) | \boldsymbol{X}_{k-1}] = -(\hat{\theta}_k - \theta_k), \quad k \in \mathbb{N}_0.$$

General Gain Construction

- Assume each measure in $\mathcal{P}_k = \{P_\theta(x|\boldsymbol{X}_{k-1}), \theta \in \Theta \subset \mathbb{R}^d\}$ has a density $p_\theta(x|\boldsymbol{X}_{k-1})$, $\theta \in \Theta$, with respect to some σ-finite dominating measure.

- Assume also that there is a common support $\mathcal{X}$ for these densities, and that for any $x \in \mathcal{X}$, $\boldsymbol{x}_{k-1} \in \mathcal{X}^{k-1}$, and $\theta \in \Theta$, the partial derivatives $\partial p_\theta(x|\boldsymbol{x}_{k-1})/\partial\theta_i$, $i = 1, \ldots, d$, exist and are finite.

- Let $\nabla_\theta \log p_\theta(x|\boldsymbol{x}_{k-1})$ be a gradient.

General Gain Construction

We can use the gains

$$G_k(\vartheta, \boldsymbol{x}_k) = \nabla_\vartheta \log p_\vartheta(x_k | \boldsymbol{x}_{k-1}).$$

If expectation and differentiation can be interchanged, then

$$\begin{aligned} g_k(\vartheta, \theta | \boldsymbol{X}_{k-1}) &= \mathbb{E}_\theta \left[\nabla_\vartheta \log p_\vartheta(X_k | \boldsymbol{X}_{k-1}) \middle| \boldsymbol{X}_{k-1} \right] \\ &= \nabla_\vartheta \mathbb{E}_\theta \left[\log p_\vartheta(X_k | \boldsymbol{X}_{k-1}) \middle| \boldsymbol{X}_{k-1} \right] \\ &= -\nabla_\vartheta KL\big(P_\theta(\cdot | \boldsymbol{X}_{k-1}), P_\vartheta(\cdot | \boldsymbol{X}_{k-1})\big). \end{aligned}$$

Motivation

- In quantile regression we estimate a conditional quantile (rather than conditional expectation) of one random variable given another.
- More robust than regression: no moment assumptions.
- Estimating multiple quantiles simultaneously gives more comprehensive picture of the distribution.
- Relevant in econometrics, social sciences and ecology.
- Relation between response variable and the measured predictors might be complex and not be captured by the conditional expectation.

Notion of Depth

- Multidimensional analogues of the median, center points of a distribution, can be defined using a depth function.
- Given $\mathbb{P}$ with support in $\mathbb{R}^d$, the depth of $x \in \mathbb{R}^d$ with respect to $\mathbb{P}$, $DF(x, \mathbb{P})$, measures the *centrality* of x in $\mathbb{P}$.
- Depth functions should give a $\mathbb{P}$-based, center-outward ordering of the points $x \in \mathbb{R}^d$ via contours of the function $x \mapsto DF(x, \mathbb{P})$.
- Points of maxima of the depth function $DF(x, \mathbb{P})$ will be the "most central" points of the distribution $\mathbb{P}$.

Properties of the Half-Space Depth

- Tukey's notion of depth, the half-space depth, is defined as

$$DF(x,\mathbb{P}) = \inf\Big\{\mathbb{P}\big(H\big) : H \text{ is a closed half-space}, x \in H\Big\}.$$

- In one dimension, points of maximum half-space depth are, by definition, medians.

- The half-space depth has attractive properties, namely:
  - it is invariant under affine transformations;
  - for distributions with a natural notion of center, it attains its maximum at this center;
  - it decays monotonically relative to its deepest point;
  - it vanishes at infinity.

Half-Space Symmetrical Distributions

Quantiles and spatial medians

- We work with distributions for which there is a proper notion of center.
- Absolutely continuous distributions $\mathbb{P}$ which are half-space symmetric about θ:

$$\theta \in H \Rightarrow \mathbb{P}(H) \geq 1/2, \; H \text{ is a closed half-space.}$$

- In one dimension, for $\alpha \in [0, 1/2)$,

$$\theta(\alpha) = \inf\{x \in \mathbb{R} : DF(x, \mathbb{P}) \geq \alpha\},$$
$$\theta(1-\alpha) = \sup\{x \in \mathbb{R} : DF(x, \mathbb{P}) \geq \alpha\},$$

- In d dimensions, $d \geq 1$,

$$\theta(1/2) = \{x \in \mathbb{R}^d : DF(x, \mathbb{P}) \geq 1/2\}.$$

Model

Assume we observe $\boldsymbol{X}_n = (X_0, X_1, \ldots, X_n)$ according to:

$$X_0 \sim \mathbb{P}_0, \qquad X_k | \boldsymbol{X}_{k-1} \sim \mathbb{P}_k(\cdot | \boldsymbol{X}_{k-1}), \quad k \in \mathbb{N},$$

where these conditional measures have support $\mathcal{X} \subset \mathbb{R}^d$.

Objective

Given $\alpha_k \in (0,1)$, $k \in \mathbb{N}$, we would like to track $\theta_k(\boldsymbol{X}_{k-1}, \alpha_k)$.

If $d = 1$ we track $\theta_k(\boldsymbol{X}_{k-1}, \alpha_k)$, the conditional quantile of level α_k of $\mathbb{P}_k(\cdot | \boldsymbol{X}_{k-1})$.

If $d \geq 2$ we fix $\alpha_k = 1/2$, $k \in \mathbb{N}$, and track $\theta_k(\boldsymbol{X}_{k-1}, 1/2)$, the conditional spatial median of $\mathbb{P}_k(\cdot | \boldsymbol{X}_{k-1})$.

Assumptions

Let $H(x,w)$ be the half-space containing $x+w$, delimited by the hyper-plane which contains x and is perpendicular to w.

Assumption (B1)

For $b, B, \delta > 0$, any $\epsilon \in (0,\delta]$, any unit vectors $v, w \in \mathbb{R}^d$:

$$b \leq \frac{\mathbb{P}_k\big(H(\theta_k - \epsilon v, w)\big|\boldsymbol{x}_{k-1}\big) - \alpha_k}{v^T w\, \epsilon} \leq B,\ \boldsymbol{x}_k \in \mathcal{X}^k,$$

where $k \in \mathbb{N}$, $\theta_k = \theta_k(\boldsymbol{x}_{k-1}, \alpha_k)$ as before.

Assumption (B2)

The support $\mathcal{X}$ is a compact set, such that for all $x \in \mathcal{X}$, $\|x\| \leq C_{\mathcal{X}}$ and the conditional spatial quantiles θ_k take values in some convex subset $\Theta \subseteq \mathcal{X}$ with $\|\theta\| \leq C_\Theta$, $\theta \in \Theta$.

Definition of the Gains and Algorithms

Proposed Gains

- For $d = 1$ $u \in \mathcal{X}$, $v \in \mathbb{R}$ and $\alpha \in (0, 1)$

$$R(u, v, \alpha) = \alpha - I\{u \leq v\}, \quad k \in \mathbb{N}.$$

- For $d \geq 2$, $u \in \mathcal{X} \subset \mathbb{R}^d$, $v \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$ a unit vector

$$S(u, v, w) = w\Big(I\{u \in H(v, w)\} - 1/2\Big), \quad k \in \mathbb{N}.$$

Let $\boldsymbol{B} = \{e_1, \ldots, e_d\}$ be an orthonormal basis for $\mathbb{R}^d$. Call D such that $\mathbb{P}(D = e_i) = 1/d$, $i = 1, \ldots, d$, a random direction.

Definition of the Gains and Algorithms

Proposed Gains

- For $d = 1$ $u \in \mathcal{X}$, $v \in \mathbb{R}$ and $\alpha \in (0,1)$
$$R(u,v,\alpha) = \alpha - I\{u \leq v\}, \quad k \in \mathbb{N}.$$
- For $d \geq 2$, $u \in \mathcal{X} \subset \mathbb{R}^d$, $v \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$ a unit vector
$$S(u,v,w) = w\Big(I\{u \in H(v,w)\} - 1/2\Big), \quad k \in \mathbb{N}.$$

Proposed Algorithms

- For $d = 1$, $\alpha_k \in (0,1)$, $k \in \mathbb{N}$,
$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k R(X_k, \hat{\theta}_k, \alpha_k), \quad \hat{\theta}_1 \in \Theta, \ k \in \mathbb{N},$$
- For $d \geq 2$ and D_k a sequence of i.i.d. random directions
$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_k S(X_k, \hat{\theta}_k, D_k), \quad \hat{\theta}_1 \in \Theta, \ k \in \mathbb{N},$$

Preliminary Results

Verifying assumption (A1) and assumption (A2)

Lemma (Representation $d = 1$)

A.s. $\|R(X_k, \hat{\theta}_k, \alpha_k)\| \leq 1$. *If (B1) and (B2) hold, then*

$$\mathbb{E}\big[R(X_k, \hat{\theta}_k, \alpha_k) \big| \mathcal{F}_{k-1}\big] = -M_k(\hat{\theta}_k - \theta_k), \quad k \in \mathbb{N},$$

with $\mathcal{F}_{k-1} = \sigma(\boldsymbol{D}_{k-1}, \boldsymbol{X}_{k-1})$, *for* $\mathcal{F}_{k-1}$*-measurable random variables* M_k *such that a.s.* $0 < \lambda_1 \leq M_k \leq \lambda_2 < \infty$.

Lemma (Representation $d \geq 2$)

A.s. $\|S(X_k, \hat{\theta}_k, D_k)\| \leq 1/2$. *If (B1) and (B2) hold, then*

$$\mathbb{E}\big[S(X_k, \hat{\theta}_k, D_k) \big| \mathcal{F}_{k-1}\big] = -\boldsymbol{M}_k(\hat{\theta}_k - \theta_k), \quad k \in \mathbb{N},$$

with $\mathcal{F}_{k-1} = \sigma(\boldsymbol{D}_{k-1}, \boldsymbol{X}_{k-1})$, *for* $\mathcal{F}_{k-1}$*-measurable matrices* $\boldsymbol{M}_k$ *such that a.s.* $0 < \lambda_1 \leq \lambda_{(1)}(\boldsymbol{M}_k) \leq \lambda_{(d)}(\boldsymbol{M}_k) \leq \lambda_2 < \infty$.

References

References

- G. Bassett and R. Koenker, **Asymptotic theory of least absolute error regression**, JASA, 73(363):618–622, 1978.
- E. Belitser and P. Serra, **On properties of the algorithm for pursuing a drifting quantile**, Automation and Remote Control, 74(4):613–627, 2013.
- E. Belitser and P. Serra, **Online Tracking of a Predictable Drifting Parameter of a Time Series**, arXiv:1306.0325 [math.ST], 2013.
- E. Belitser and P. Serra, **Tracking of a Conditional Spatial Median**, pre-print, 2013.
- R. Koenker, **Quantile Regression**, Cambridge University Press, 2005.
- H. J. Kushner and G. Yin, **Stochastic Approximation and Recursive Algorithms and Applications**, Berlin and New York: Springer-Verlag, 2003.
- H. Robbins and S. Monro, **A stochastic approximation method**, The Annals of Mathematical Statistics, 22(3):400–407, 1951.
- J. W. Tukey, **Mathematics and the picturing of data**, Proceedings of the international congress of mathematicians, volume 2, pages 523–531, 1975.
- M. T. Wasan, **Stochastic Approximation**, Cambridge University Press, 1969.
- Y. Zuo and R. Serfling, **General notions of statistical depth function**, Annals of Statistics, pages 461–482, 2000.