FMS051

Take Home Exam 2022

Arvid Gramer

I have not collaborated with anyone, apart from questions to the teaching staff.

Arvid Gramer Lund 22-01-10

4) Consider a real-valued stationary process x_t with mean m_x. Show that the optimal predictor $\hat{x}_{t+k|t}$ of form $ax_t + b$ is obtained by choosing $a = \rho_x(k)$ and $b = m_x - m_x \rho_x(k)$, where $\rho_x(k)$ denotes autocorrelation of x_t.

$$\rho_x(k) = \frac{r_x(k)}{r_x(0)}.$$

The optimal predictor is found as the projection of x_{t+k} onto space spanned by x_t, namely

$$\hat{x}_{t+k|t} = E[x_{t+k} | x_t] = a x_t + b.$$

We want to minimize the prediction error variance:

$$V[\hat{x}_{t+k|t} - x_{t+k}] = V[a x_t + b - x_{t+k}]$$

This can be written as

$$V[a x_t + b - x_{t+k}] = C[ax_t, ax_t] + C[ax_t, b] + C[ax_t, -x_{t+k}]$$
$$+ C[b, ax_t] + C[b, b] + C[b, -x_{t+k}] + C[-x_{t+k}, ax_t]$$
$$+ C[-x_{t+k}, b] + C[-x_{t+k}, -x_{t+k}].$$

Since b is constant, all terms with $b = 0$.
Using $C[x_t, x_{t+k}] = r_x(k)$:

$$a^2 r_x(0) - a r_x(k) - a r_x(-k) + r_x(0).$$

with stationarity $r_x(-k) = r_x^*(k)$. Finding minima using

$$\frac{\partial V}{\partial a} = 2a r_x(0) - 2 r_x(k) = 0 \Rightarrow \quad a = \frac{r_x(k)}{r_x(0)} = \rho_x(k)$$

Arvid Gramer

1: Furthermore we want an unbiased estimator, namely:

$$E[\hat{X}_{t+k|t} - X_{t+k}] = 0$$

$$\Rightarrow \quad E[aX_t + b - X_{t+k}] = aE[X_t] + E[b] - E[X_{t+k}] = 0$$

with stationarity $E[X_t] = m_X \quad \forall t$:

$$a m_X + b - m_X = 0 \quad \Rightarrow \quad b = m_X - a m_X$$

This yields:

$$\begin{cases} a = \rho_X(k) \\ b = m_X - \rho_X(k) m_X \end{cases}$$

Arvid Gramer

2: Let $\{X_t\}$ denote a (real-valued) sequence of independent normal random variables, each with zero mean and variance σ_X^2. Determine values of the constant c such that

$$Y_t = \sin(ct) X_t + \cos(ct) X_{t-1}$$

is a weekly stationary process.

From page 39 and 42 in the course book we have some conditions that need to be fulfilled for a weakly stationary process.

To begin with, we need a constant and finite mean.

$$E[Y_t] = E[\sin(ct) X_t] + E[\cos(ct) X_{t-1}]$$

$$= E[\sin(ct)] E[X_t] + E[\cos(ct)] E[X_{t-1}] = 0$$

Furthermore, we want the autocovariance $C[Y_t, Y_{t+\tau}]$ to only depend on τ, $C[Y_t, Y_{t+\tau}] = \gamma_Y(\tau)$.

This becomes:

$$C[Y_t, Y_{t+\tau}] = C[\sin(ct) X_t + \cos(ct) X_{t-1}, \sin(c(t+\tau)) X_{t+\tau} + \cos(c(t+\tau)) X_{t+\tau-1}]$$

$$\Rightarrow = C[\sin(ct) X_t, \sin(c(t+\tau)) X_{t+\tau}] + C[\sin(ct) X_t, \cos(c(t+\tau)) X_{t+\tau-1}]$$

$$+ C[\cos(ct) X_{t-1}, \sin(c(t+\tau)) X_{t+\tau}] + C[\cos(ct) X_{t-1}, \cos(c(t+\tau)) X_{t+\tau-1}]$$

Arvid Gramer

2 Since the trigonometric functions are deterministic we have for example

Arvid Gramer

$$C[\sin(ct)\, x_t,\ \sin(c(t+\tau))\, x_{t+\tau}] = E[\sin(ct)\sin(c(t+\tau))\, x_t\, x_{t+\tau}]$$

$$= \sin(ct)\sin(c(t+\tau))\, E[x_t\, x_{t+\tau}] = \sin(ct)\sin(c(t+\tau))\, C[x_t, x_{t+\tau}]$$

This holds for all terms, yielding

$$\sin(ct)\sin(c(t+\tau))\, r_x(\tau) + \sin(ct)\cos(c(t+\tau))\, r_x(\tau-1)$$
$$+ \cos(ct)\sin(c(t+\tau))\, r_x(\tau+1) + \cos(ct)\cos(c(t+\tau))\, r_x(\tau)$$

We now have an expression for $r_y(\tau)$ that needs to fulfill

$$r_y(\tau) = r_y^*(-\tau) \quad (= r_y(-\tau) \text{ due to real-valued})$$

$$r_y(0) \geq \begin{cases} 0 \\ |r_y(\tau)| \ \forall \tau \end{cases}$$

For $\tau = 0$ we have (since $r_x(k) = \begin{cases} \sigma_x^2 & k=0 \\ 0 & \text{else} \end{cases}$)

$$r_y(0) = \sin^2(ct)\,\sigma_x^2 + \cos^2(ct)\,\sigma_x^2 = \sigma_x^2 > 0 \quad :)$$

$$r_y(-1) = \cos(ct)\sin(c(t-1))\,\sigma_x^2 \leq \sigma_x^2 \quad :)$$

$$r_y(1) = \sin(ct)\cos(c(t+1))\,\sigma_x^2 \leq \sigma_x^2 \quad :)$$

$$r_y(n) = 0 \quad |n| > 1.$$

We also need $r_y(-1) = r_y(1)$:

$$\cos(ct)\sin(ct-c) \overset{!}{=} \sin(ct)\cos(ct+c)$$

2: Using trigonometric identities : $\begin{cases} \sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta \\ \cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{cases}$

$$\Rightarrow \cos(ct)\left(\cancel{\sin(ct)}\cos(c) - \cos(ct)\sin(c)\right) \stackrel{!}{=} \sin(ct)\left(\cancel{\cos(ct)}\cos(c) - \sin(ct)\sin(c)\right)$$

$$\Leftrightarrow \left(\cos^2(ct) - \sin^2(ct)\right)\sin(c) = 0$$

$$\Rightarrow \cos(2ct)\sin(c) = 0 \;.\; \Rightarrow \begin{cases} c = k\pi \;, \quad k = 0, \pm 1, \pm 2 \ldots \\ 2ct = \frac{\pi}{2} + m\pi \Rightarrow c = \frac{\pi(\frac{1}{2}+m)}{2t} \end{cases}$$

not valid for all t.
Also should not depend on t

Answer $\underline{c = k\pi}$, $k \in \mathbb{Z}$

Arvid Gramer

Arvid Graner

3 Let

$$x_t + a_1 x_{t-1} + a_2 x_{t-2} = e_t$$

e_t white noise, zero mean, variance σ_e^2.

AR(2)-process.

a) Derive 1- and 2-step predictors for x_t.

From theorem 6.6 we have that the optimal linear predictor of x_{t+k} is

$$\hat{x}_{t+k|t} = \frac{G(z)}{C(z)} x_t$$

For us, $C(z) = 1$ and we get $G(z)$ from the Diophantine equation

$$C(z) = A(z)F(z) + z^{-k}G(z).$$

For us
$$\begin{cases} A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} \\ \operatorname{ord}(F(z)) = k-1 = 0 \text{ or } 1 \\ \operatorname{ord}(G(z)) = \max(p-1, q-k) = 1. \end{cases}$$

This yields for $k=1$, $F(z) = 1$. (monic, order 0)

$$1 = (1 + a_1 z^{-1} + a_2 z^{-2}) \quad z^{-1}(g_0 + g_1 z^{-1})$$

$$\Rightarrow (a_1 + g_0) z^{-1} + (a_2 + g_1) z^{-2} = 0 \Rightarrow \begin{cases} g_0 = -a_1 \\ g_1 = -a_2 \end{cases}$$

Could have guessed since it is an AR.

For the $k=2$ step prediction, we solve the equation: $\mathrm{Ord}(F(z)) = 1$.

Arvid Gramer

$$1 = (1 + a_1 z^{-1} + a_2 z^{-2})(1 + f_1 z^{-1}) + z^{-2}(g_0 + g_1 z^{-1})$$

$$\Rightarrow \quad 1 = 1 + f_1 z^{-1} + a_1 z^{-1} + a_1 f_1 z^{-2} + a_2 z^{-2} + a_2 f_1 z^{-3} + g_0 z^{-2} + g_1 z^{-3}$$

$$\Rightarrow (f_1 + a_1) z^{-1} + (a_1 f_1 + a_2 + g_0) z^{-2} + (a_2 f_1 + g_1) z^{-3} = 0$$

$$f_1 = -a_1 \qquad g_0 = a_1^2 - a_2 \qquad g_1 = a_1 a_2$$

Our one step prediction becomes

$$\hat{x}_{t+1|t} = (-a_1 - a_2 z^{-1}) x_t = -a_1 x_t - a_2 x_{t-1}$$

$$\hat{x}_{t+2|t} = (a_1^2 - a_2 + a_1 a_2 z^{-1}) x_t = (a_1^2 - a_2) x_t + a_1 a_2 x_{t-1}$$

This is fairly unsurprising since it is an AR-process we could just "move" the terms to the other side k times.

$$E[\hat{x}_{t+1|t} + a_1 x_t + a_2 x_{t-1}] = E[e_t] = 0$$

At time t x_t and x_{t-1} known.

$$\Rightarrow E[\hat{x}_{t+1|t}] = E[-a_1 x_t - a_2 x_{t-1}] = -a_1 x_t - a_2 x_{t-1}$$

The bonus of our solution is that the error variance is found instantly as (6.45 in book)

$$V[\varepsilon_{t+k|t}] = (1 + f_1^2 + \dots + f_{k-1}^2)\sigma_e^2$$

Yielding one step error variance $V[\varepsilon_{t+1|t}] = \sigma_e^2$

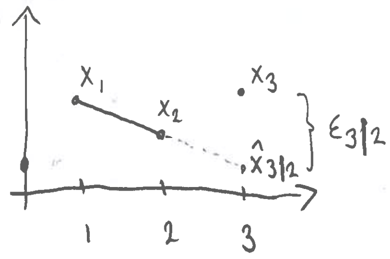
two step $V[\varepsilon_{t+2|t}] = (1 + a_1^2)\sigma_e^2$

3b) The extrapolation predictor means that our prediction at $t+k$ is the line between X_{t-1} and X_t prolonged k steps.

Arvid Gramer

This is formed as $\frac{\Delta X}{\Delta t}$

$$\hat{X}_{t+k|t} = X_t + \frac{X_t - X_{t-1}}{\underbrace{t-(t-1)}_{=1}} \cdot k$$



Our prediction error is then:

$$\varepsilon_{t+k|t} = \hat{X}_{t+k|t} - X_{t+k} = (X_t + kX_t - kX_{t-1}) - X_{t+k}$$

The variance becomes:

$$V[\varepsilon_{t+k|t}] = C[(1+k)X_t - kX_{t-1} - X_{t+k},\ (1+k)X_t - kX_{t-1} - X_{t+k}]$$

$$= (1+k)^2 r_X(0) - k(1+k)r_X(-1) - (1+k)r_X(k)$$

$$-k(1+k)r_X(1) + k^2 r_X(0) + k\, r_X(k+1)$$

$$+ r_X(0) - (1+k)r_X(-k) + k\, r_X(-(k+1))$$

For these covariances we need to solve Yule-Walker equations. We need $r_X(\tau)$ for τ up to $k+1 = 3$. Begin with $\tau = 0, 1, 2$.

$$r_X(0) + a_1 r_X(1) + a_2 r_X(2) = \sigma_e^2$$

$$r_X(1) + a_1 r_X(0) + a_2 r_X(1) = 0$$

$$r_X(2) + a_1 r_X(1) + a_2 r_X(0) = 0$$

Yields:

$$\begin{bmatrix} 1 & a_1 & a_2 \\ a_1 & 1+a_2 & 0 \\ a_2 & a_1 & 1 \end{bmatrix} \begin{bmatrix} r_X(0) \\ r_X(1) \\ r_X(2) \end{bmatrix} = \begin{bmatrix} \sigma_e^2 \\ 0 \\ 0 \end{bmatrix}$$

Alvid Gramer

3b) This is a quite cumbersome task.
Luckily, there is a explicit solution in "Stationary stochastic processes for scientists and Engineers" by Lindgren, Rootzén and Sandsten on page 171, (7.6):

Standing on the shoulders of giants.

$$\begin{cases} r_x(0) = \dfrac{\sigma_e^2}{(1+a_2)^2 - a_1^2} \dfrac{1+a_2}{1-a_2} \\ r_x(1) = -\dfrac{a_1}{1+a_2} \cdot r(0) := Q_1 r(0) \\ r_x(2) = \dfrac{a_1^2 - a_2 + a_2^2}{1+a_2} \cdot r(0) := Q_2 r(0) \end{cases}$$

For $r(3)$ we use.

$$r_x(3) + a_1 r(2) + a_2 r(1) = 0$$

$$\Rightarrow r_x(3) + \left(\frac{a_1(a_1^2 - a_2 + a_2^2)}{1+a_2} + \frac{a_2(-a_1)}{1+a_2} \right) r(0) = 0$$

$$\Rightarrow r_x(3) = -\frac{a_1(a_1^2 + a_2^2)}{1+a_2} r(0) := Q_3 r(0)$$

From $V[\varepsilon_{t+k|t}]$ we get: (using $r(-\tau) = r(\tau)$)

$$V[\varepsilon_{t+k|t}] = ((1+k)^2 + k^2 + 1) r_x(0) + (-2k(1+k)) r_x(1) + 2k r_x(k+1) - 2(1+k) r_x(k)$$

For 1-step prediction:

$$V[\varepsilon_{t+1|t}] = 6 r_x(0) - 8 r_x(1) + 2 r_x(2)$$

$$= 6 r_x(0) - 8 Q_1 r_x(0) + 2 Q_2 r_x(0)$$

$$= (3 - 4Q_1 + Q_2) 2 r_x(0)$$

3b For 2-step prediction: Arvid Gramer

$$V[\varepsilon_{t+2|t}] = 14\, r_x(0) - 12\, r_x(1) - 6\, r_x(2) + 4\, r_x(3)$$

$$= 14\, r_x(0) - 12\, Q_1 r_x(0) - 6\, Q_2 r_x(0) + 4\, Q_3 r_x(0)$$

Answer: The error variances for the one and two step extrapolator predictors are

$$V[\varepsilon_{t+1|t}] = 2(3 - 4Q_1 + Q_2)\, r_x(0)$$

$$V[\varepsilon_{t+2|t}] = 2(7 - 6Q_1 - 3Q_2 + 4Q_3)\, r_x(0)$$

where

$$\begin{cases} r_x(0) = \dfrac{\sigma_e^2}{(1+a_2)^2 - a_1^2} \dfrac{1+a_2}{1-a_2} \\ Q_1 = \dfrac{-a_1}{1+a_2} \\ Q_2 = \dfrac{a_1^2 - a_2 + a_2^2}{1+a_2} \\ Q_3 = \dfrac{-a_1(a_1^2 + a_2^2)}{1+a_2} \end{cases}$$

Arvid Bramer

3c) Determine a_1 and a_2 for which the extrapolating estimator optimal, if it exists.

The extrapolating predictor gives a prediction

$$\hat{X}_{t+k|t} = X_t + k(X_t - X_{t-1}) = (k+1)X_t - kX_{t-1}$$

For this to be an optimal predictor we need to have a process where the difference between each time step t is constant, + same noise.

This would mean that

$$X_{t+k} = \hat{X}_{t+k|t} + e_t = (k+1)X_t - kX_{t-1} + e_t$$

This gives, for $k=1$:

$$X_{t+1} = 2X_t - X_{t-1} + e_t$$

This translates to a process

$$X_t - 2X_{t-1} + X_{t-2} = e_t \Rightarrow \begin{cases} a_1 = -2 \\ a_2 = 1 \end{cases}$$

A way to verify this is that our optimal linear predictor in 3a) is

$$\hat{X}_{t+k} = \begin{cases} -a_1 X_t - a_2 X_{t-1} & k=1 \\ (a_1^2 - a_2)X_t + a_1 a_2 X_{t-1} & k=2 \end{cases}$$

Identifying coefficients yield:

$$-a_1 = 2 \quad , \quad -a_2 = -1 \quad .$$

This also fulfills the two step $\hat{X}_{t+2} = 3X_t - 2X$ since

$$\begin{cases} a_1^2 - a_2 = 4 - 1 = 3 \\ a_1 a_2 = -2 \cdot 1 = -2 \end{cases}$$

Arvid Gramer

However this process would not be stable, since it would grow linearly and not be bounded. [$A(z)$ having roots on the unit circle]
It would also mess up our variance since $a_2 = 1$ makes our expression for $r_x(0)$ to be singular.

This leads us to the conclusion that it is not an optimal predictor for any w.s.s process.

Arvid Gramer

4. We have a linear process (real valued)

$$y_t = x_t^T \theta + e_t$$

We want to derive a way to update our parameters θ recursively such that when new information y_t, x_t is available, we do not need to recalculate everything, more tweak the current parameters.

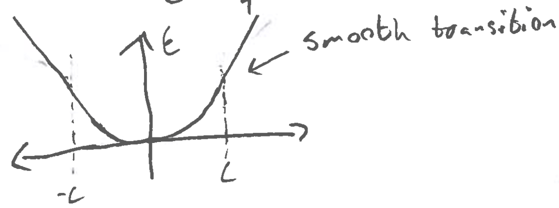
The cost function we want to minimize is

$$f(t, \theta) = \sum_{\ell=1}^{t} \rho\left(\sigma_e^{-1}(y_\ell - x_\ell^T \theta)\right) \sigma_e^2$$

where

$$\rho(u) = \begin{cases} \frac{u^2}{2} & |u| \leq c \\ c|u| - \frac{c^2}{2} & |u| > c \end{cases}$$

which is then linear when $|u| > c$ and quadratic when $|u| \leq c$. Something like



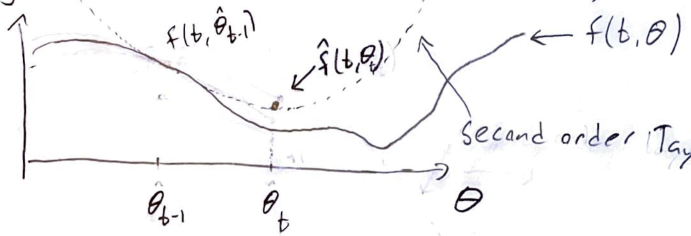
Using the approximation

$$f(t, \theta) \approx f(t, \hat{\theta}_{t-1}) + (\theta - \hat{\theta}_{t-1})^T f'(t, \hat{\theta}_{t-1})$$
$$+ \frac{1}{2}(\theta - \hat{\theta}_{t-1})^T f''(t, \hat{\theta}_{t-1})(\theta - \hat{\theta}_{t-1})$$

~~the~~ means that we want to find new parameters θ which minimizes the new cost $f(t, \theta)$ approximated as the Taylor expansion around the cost for the old parameters $\hat{\theta}_{t-1}$.

Arvid Gramer

Meaning that (shown for one dimensional θ:)



$\hat{f}(t,\theta_t)$ is the approximation of $f(t,\theta_t)$

This means that our best guess of how the cost function behaves is to use the parameter estimate from $t-1$ and add the cost generated by the latest observation Y_t, X_t, creating $f(t,\hat{\theta}_{t-1})$. We then approximate the proximity of this point as the second order Taylor, and use the parameters θ that minimizes this value. This becomes

$$\hat{f}(t,\theta) = f(t,\hat{\theta}_{t-1}) + \overbrace{\begin{bmatrix}\theta_1-\hat{\theta}_1 & \theta_2-\hat{\theta}_2 & \cdots & \theta_n-\hat{\theta}_n\end{bmatrix}}^{\text{at } t-1} \begin{bmatrix} f'_{\theta_1}(t,\hat{\theta}_{t-1}) \\ f'_{\theta_2}(t,\hat{\theta}_{t-1}) \\ \vdots \\ f'_{\theta_n}(t,\hat{\theta}_{t-1}) \end{bmatrix}$$

$$+\frac{1}{2}\begin{bmatrix}\theta_1-\hat{\theta}_1 & \theta_2-\hat{\theta}_2 & \theta_n-\hat{\theta}_n\end{bmatrix} \begin{bmatrix} f''_{\theta_1\theta_1} & f''_{\theta_1\theta_2} & \cdots & f''_{\theta_1\theta_n} \\ f''_{\theta_2\theta_1} & f''_{\theta_2\theta_2} & & \vdots \\ f''_{\theta_n\theta_1} & \cdots & & f''_{\theta_n\theta_n} \end{bmatrix} \begin{bmatrix} \theta_1-\hat{\theta}_1 \\ \theta_2-\hat{\theta}_2 \\ \vdots \\ \hat{\theta}_n-\theta_n \end{bmatrix}$$

In the vectors and matrices the indices represent position and not time
n is the number of parameters.

$$\hat{f}(t,\theta) = f(t,\hat{\theta}_{t-1}) + \sum_{i=1}^{n}(\theta_i-\hat{\theta}_{i|t-1})\, f'_{\theta_i}(t,\hat{\theta}_{i|t-1})$$

$$+\frac{1}{2}\sum_{j=1}^{n}(\theta_j-\hat{\theta}_{j|t-1})\left(\sum_{k=1}^{n} f''_{\theta_j\theta_k}(\theta_k-\hat{\theta}_{k|t-1})\right)$$

Optimizing wrt θ is now fairly simple. Arvid Gramer

We are looking for the θ that minimizes our expression. We do this by derivating wrt θ and finding where this is zero. $f(t, \hat{\theta}_{t-1})$, $f'(t, \hat{\theta}_{t-1})$ and $f''(t, \hat{\theta}_{t-1})$ are all previous values and do not change when we vary θ: (Omitting index $t-1$)

$$\begin{bmatrix} \frac{\partial \hat{f}}{\partial \theta_1} \\ \frac{\partial \hat{f}}{\partial \theta_2} \\ \vdots \\ \frac{\partial \hat{f}}{\partial \theta_n} \end{bmatrix} = \begin{bmatrix} f'_{\theta_1}(t, \hat{\theta}) \\ f'_{\theta_2}(t, \hat{\theta}) \\ \vdots \\ f'_{\theta_n}(t, \hat{\theta}) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \sum_{k=1}^{n} f''_{1k}(\theta_k - \hat{\theta}_k) + \sum_{j=1}^{n} (\theta_j - \hat{\theta}_j) f''_{j1} \\ \sum_{k=1}^{n} f''_{2k}(\theta_k - \hat{\theta}_k) + \sum_{j=1}^{n} (\theta_j - \hat{\theta}_j) f''_{j2} \\ \vdots \\ \sum_{k=1}^{n} f''_{nk}(\theta_k - \hat{\theta}_k) + \sum_{j=1}^{n} (\theta_j - \hat{\theta}_j) f''_{jn} \end{bmatrix}$$

Using that $f''_{j1} = f''_{1j}$

$$\begin{bmatrix} \frac{\partial \hat{f}}{\partial \theta_1} \\ \frac{\partial \hat{f}}{\partial \theta_2} \\ \vdots \\ \frac{\partial \hat{f}}{\partial \theta_n} \end{bmatrix} = \begin{bmatrix} f'_{\theta_1}(t, \hat{\theta}) \\ f'_{\theta_2}(t, \hat{\theta}) \\ \vdots \\ f'_{\theta_n}(t, \hat{\theta}) \end{bmatrix} + \begin{bmatrix} \sum_{k=1}^{n} f''_{1k}(\theta_k - \hat{\theta}_k) \\ \sum_{k=1}^{n} f''_{2k}(\theta_k - \hat{\theta}_k) \\ \vdots \\ \sum_{k=1}^{n} f''_{nk}(\theta_k - \hat{\theta}_k) \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We therefore want to solve

$$f'_{\theta_i}(t, \hat{\theta}_{t-1}) + \sum_{k=1}^{n} f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_k) = 0$$

For all $\theta_i \dots \theta_n$.

For the one dimensional case this becomes

$$f'(t, \hat{\theta}_{t-1}) + f''\theta - f''\hat{\theta}_{t-1} = 0$$

$$\Rightarrow \quad \theta = \hat{\theta}_{t-1} - \frac{f'(t, \hat{\theta}_{t-1})}{f''(t, \hat{\theta}_{t-1})}$$

We now only need f' and f''

Derivating the expression: Arvid Gramer

$$f(t,\theta) = \sum_{\ell=1}^{t} \rho(\sigma_e^{-1}(y_\ell - x_\ell^T \hat{\theta}_{t-1}))\sigma_e^2$$

$$\frac{\partial}{\partial\theta_i} f = \sum_{\ell=1}^{t} \frac{\partial}{\partial\theta_i} \rho(\sigma_e^{-1}(y_\ell - x_\ell^T \hat{\theta}_{t-1}))\sigma_e^2$$

Since the argument of ρ is negative linear our inner derivative is only $-x_i\sigma_e^{-1}$. With $u = \sigma_e^{-1}(y_\ell - x_\ell^T \hat{\theta}_{t-1})$

$$\frac{\partial\rho}{\partial\theta} = \frac{\partial u}{\partial\theta}\frac{\partial\rho}{\partial u} = -\sigma_e^{-1}x_i \begin{cases} u & \text{if } |u| < c \\ c & \text{if } u \geq c \\ -c & \text{if } u \leq c \end{cases}$$

$$= -x\sigma_e^{-1} \begin{cases} \sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1}) & \text{if } |\sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1})| < c \\ c & \text{if } \sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1}) \geq c \\ -c & \text{if } \sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1}) \leq c \end{cases}$$

Second derivative:

$$\frac{\partial}{\partial\theta}\left(\frac{\partial\rho}{\partial\theta}\right) = \frac{\partial u}{\partial\theta}\frac{\partial}{\partial u}\left(\frac{\partial\rho}{\partial\theta}\right) = \begin{cases} x^2\sigma_e^{-2} & \text{if } |\sigma^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1})| < c \\ 0 & \text{if } |\sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1})| \geq c \end{cases}$$

This gives:

$$\frac{\partial f}{\partial\theta_i} = \sum_{\ell=1}^{t} \begin{cases} -x_{i\ell}(y_\ell - x_\ell^T\hat{\theta}_{t-1}) & \text{if } |\sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}_{t-1})| < c \\ -x_i c\sigma_e & \text{if } \sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}) \geq c \\ x_i c\sigma_e & \text{if } \sigma_e^{-1}(y_\ell - x_\ell^T\hat{\theta}) \leq c \end{cases}$$

And

Arvid Gran

$$\frac{\partial^2 f}{\partial \theta_i \theta_j} = \frac{\partial^2 f}{\partial \theta_j \theta_i} = \sum_{t=1}^{t} \begin{cases} x_i x_j \sigma_e^{-1} & \text{if } |\sigma_e^{-1}(y_t - x_t^T \theta)| < c \\ 0 & \text{else} \end{cases}$$

Now this is a slight problem, since the expression for our new θ:

$$f'_{\theta_i}(t, \hat{\theta}_{t-1}) + \sum_{k=1}^{n} f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_{k(t-1)}) = 0$$

$$f'_{\theta_i}(t, \hat{\theta}_{t-1}) + f''_{\theta_i \theta_i}(\theta_i - \hat{\theta}_i) + \sum_{k=1}^{i-1} f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_{k(t-1)}) + \sum_{k=i+1}^{n} f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_k) = 0$$

$$\Rightarrow f''_{\theta_i \theta_i} \theta_i = f''_{\theta_i \theta_i} \hat{\theta}_i - f'_{\theta_i} - \sum_{k \neq i}^{n} f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_{k(t-1)})$$

$$\Rightarrow \theta_i = \hat{\theta}_{i(t-1)} - \frac{f'_{\theta_i}}{f''_{\theta_i \theta_i}} - \sum_{k \neq i}^{n} \frac{f''_{\theta_i \theta_k}(\theta_k - \hat{\theta}_{k(t-1)})}{f''_{\theta_i \theta_i}}$$

we could, if we are unlucky, have a second derivative that is zero. However, that would require all terms in the sum above to be 0, meaning that all our estimates would have to be very off.

To conclude and specify our algorithm we would begin with a first model estimate θ. This should be a good guess of the true model, maybe a previous non-recursive model, so that the error $y_1 - x_1^T \theta$ is not to big so that the second derivative of the Huber cost is zero. We put this in the cost function $f(1, \theta_1)$.

We then, for all time steps t, take our previous parameter estimate θ_{t-1}, and make a prediction $x_t^T \hat{\theta}_{t-1}$. The error, $y_t - x_t^T \hat{\theta}_{t-1}$ is then put in to our approximation of the new cost function. We then pick a new parameter estimate θ as the one that minimizes that approximation, namely the $\theta = \hat{\theta}_{t-1} - \frac{f'(t, \hat{\theta}_{t-1})}{f''(t, \hat{\theta}_{t-1})}$ for one dimensinal case.

Arvid Gramer

This new parameter estimate is then used for the next prediction, and next update and so on.

This "robust" algorithm differs from the ordinary RLS by making use of this Huber cost, which since it is linear for large deviations does not "blow up" as much as a quadratic error would do. That makes for a "calmer" response to errors, which could be caused by noise. It also does not include the kalman gain.