

Modelování a aproximace kovariancí v asimilaci dat

K.Eben, I. Kasanický, J.Mandel, P.Juruš

Inst.of Computer Science
Academy of Sciences of the Czech Rep., Prague

20. ledna 2014

Asimilace dat do numerických modelů počasí

- jde o filtrační problém
- update filtru (v meteorologii tzv. analýza) dá se chápat jako statistická regresní úloha nebo optimalizační úloha
- je třeba modelovat nejistotu v modelu a v datech
- pro účely tohoto příspěvku se omezíme na lineární model a normální rozdělení
- Ensemblový Kalmanův filtr (EnKF)

Update Kalmanova filtru - regresní interpretace

- \mathbf{X}^t - skutečná hodnota stavového vektoru
- \mathbf{X}^f - předpověď z minulého časového kroku
- y - vektor pozorování
- \mathbf{H} - observační operátor ($\mathbf{H}\mathbf{X}^t$ je modelový protějšek y)

$$\begin{bmatrix} \mathbf{X}^f \\ y \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{H} \end{bmatrix} \mathbf{X}^t + \begin{bmatrix} e_x \\ e_y \end{bmatrix}$$

kde chyby e_x a e_y jsou nezávislé s nulovou stř. hodnotou a mají varianční matici

$$\text{cov}\left(\begin{bmatrix} e_x \\ e_y \end{bmatrix}\right) = \begin{bmatrix} \mathbf{P} & 0 \\ 0 & \mathbf{R} \end{bmatrix}$$

Rezidua $y - \mathbf{H}\mathbf{X}^f$ se obvykle nazývají inovace.

Matice \mathbf{R} zahrnuje chybu aparatury, reprezentativity pozorování a další (retrieval atd.). Předpokládejme, že je známa.

Vynásobením obou stran rovnice maticí $\left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\mathbf{H} & \mathbf{I} \end{array} \right]$ dostaneme

$$\left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\mathbf{H} & \mathbf{I} \end{array} \right] \begin{bmatrix} \mathbf{X}^f \\ y \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\mathbf{H} & \mathbf{I} \end{array} \right] \begin{bmatrix} \mathbf{I} \\ \mathbf{H} \end{bmatrix} \mathbf{X}^t + \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\mathbf{H} & \mathbf{I} \end{array} \right] \begin{bmatrix} e_x \\ e_y \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}^f \\ y - \mathbf{H}\mathbf{X}^f \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \mathbf{X}^t + \begin{bmatrix} e_x \\ -\mathbf{H}e_x + e_y \end{bmatrix}$$

a

$$\begin{pmatrix} \mathbf{X}^f - \mathbf{X}^t \\ y - \mathbf{H}\mathbf{X}^f \end{pmatrix} \sim N(0, \left[\begin{array}{c|c} \mathbf{P} & -\mathbf{P}\mathbf{H}^T \\ \hline -\mathbf{H}\mathbf{P} & \mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R} \end{array} \right])$$

takže z vlastností subvektorů normálního rozdělení dostaneme optimální predikci jako podmíněnou střední hodnotu

$$E(\mathbf{X}^f - \mathbf{X}^t | y - \mathbf{H}\mathbf{X}^f) = -\mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}(y - \mathbf{H}\mathbf{X}^f)$$

což vede na známou rovnici pro update stavu Kalmanova filtru

$$\mathbf{X}^a \equiv \hat{\mathbf{X}}^t = \mathbf{X}^f + \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}(y - \mathbf{H}\mathbf{X}^f).$$

Dále už vypustíme index f .
Máme tedy

$$\mathbf{X}^a = \mathbf{X} + \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{X}) = \mathbf{X} + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{X}),$$

kde $\mathbf{K} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}$ je Kalmanovo zesílení (Kalman gain).

\mathbf{X} a \mathbf{y} jsou nezávislé, takže

$$\mathbf{P}^a = \text{cov}(\mathbf{X}^a) = \text{cov}((\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{X} + \mathbf{K}\mathbf{y}) = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}(\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{R}\mathbf{K}^T$$

což se dá dále napsat jako

$$\mathbf{P}^a = \mathbf{P} - \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P} = \mathbf{P}(\mathbf{I} - \mathbf{K}\mathbf{H})^T$$

Ensemblový Kalmanův filtr - kovarianční matice

Dimenze stavu je řádově 10^6 a s maticí \mathbf{P} není možno pracovat běžným způsobem.

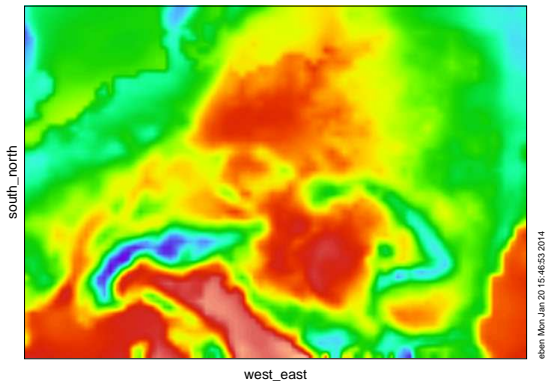
Rozdělení stavového vektoru je reprezentováno ensemblem

$\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ a kovarianční matice \mathbf{P} výběrovou kovarianční maticí

$$\hat{\mathbf{P}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N-1} \mathbb{X} (\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}^T) \mathbb{X}^T = \mathbb{Z}\mathbb{Z}^T$$

a cílem je nyní provést update stavu (tj. asimilovat data) tak, aby aktualizovaný ensemble byl rozumnou reprezentací kovarianční matice \mathbf{P}^a .

T2 (K)



Range of T2: 277.21 to 298.602 K

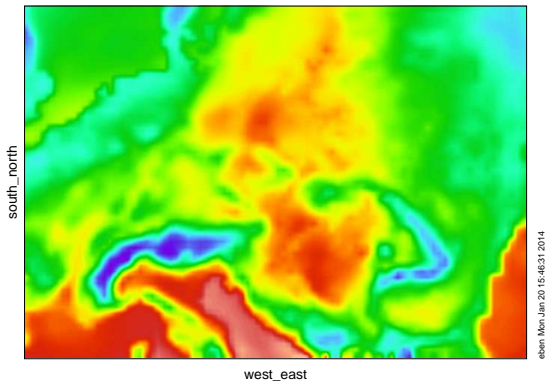
Range of west_east: 1.14301 to 31.857

Range of south_north: 42.452 to 56.5553

Current Time: 8

Frame 1 in File wrfout_d01_2012-09-11_20.00.00.e009

T2 (K)



Range of T2: 277.21 to 298.602 K
Range of west_east: 1.14301 to 31.857
Range of south_north: 42.452 to 56.5553
Current Time: 6
Frame 1 in File wrfout_d01_2012-09-11_20.00.00.e007

Ensemblový Kalmanův filtr - update step

Oprava stavového vektoru s výběrovou kovarianční maticí na místě teoretické

$$\begin{aligned}\mathbf{X}_i^a &= \mathbf{X}_i + \hat{\mathbf{P}}\mathbf{H}^T \left(\mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T + \mathbf{R} \right)^{-1} (y - \mathbf{H}\mathbf{X}_i) \\ &= \mathbf{X}_i + \mathbf{Z}\mathbf{Z}^T\mathbf{H}^T \left(\mathbf{H}\mathbf{Z}\mathbf{Z}^T\mathbf{H}^T + \mathbf{R} \right)^{-1} (y - \mathbf{H}\mathbf{X}_i)\end{aligned}$$

a průměr $\bar{\mathbf{X}}^a$ členů \mathbf{X}_i^a , tj. ensemble mean se obvykle bere jako nejlepší odhad skutečného stavu.

V praxi tedy pracujeme s maticemi $\mathbf{H}\mathbf{Z}$ s dimenzemi
(počet pozorování) x (rozsah ensemblu)
což se dá realizovat v počítači

Výběrová kovarianční matice opraveného ensamble bude

$$\hat{\mathbf{P}}^a = \frac{1}{N-1} \mathbb{X}^a (\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}^T) \mathbb{X}^{aT}$$

Jestliže ale použijeme totéž pozorování y pro update všech členů ensamble, vymizí při centrování člen s y :

$$\left(\mathbb{X} + \hat{\mathbf{K}}((y, \dots, y) - \mathbf{H}\mathbb{X}) \right) (\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}^T) = (\mathbb{X} - \hat{\mathbf{K}}\mathbf{H}\mathbb{X}) (\mathbf{I} - \frac{1}{N} \mathbf{1}\mathbf{1}^T),$$

tedy

$$\hat{\mathbf{P}}^a = (\mathbf{I} - \hat{\mathbf{K}}\mathbf{H}) \hat{\mathbf{P}} (\mathbf{I} - \hat{\mathbf{K}}\mathbf{H})^T$$

a odhad členu $\mathbf{K}\mathbf{R}\mathbf{K}^T$ chybí, tj. není postižena reprezentativita a chyby pozorování.

Update ensemblového filtru - implementace

Výše uvedený problém se řeší v zásadě dvěma způsoby

- použití perturbovaných pozorování - stochastic update
- deterministic update - square-root filtry

Další problémy vznikají v důsledku malé hodnoty výběrové kovarianční matice

- artefaktní korelace (spurious correlations)
- řešení pomocí lokalizace kovariancí, tj. použijeme Schurův součin s vhodnou korelační maticí, která tak potlačí korelace geograficky vzdálených míst

Square-root filtry

Jedna z verzí: update ensamble se rozdělí na výše uvedený update průměru (ensemble mean) a update centrovaných perturbací \mathbb{Z} . Update perturbací se provede tak, aby se dostala výběrová analogie tvaru $\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}$.

Označme tedy

$\mathbf{D} = \mathbf{H}\mathbb{Z}\mathbb{Z}^T\mathbf{H}^T + \mathbf{R}$, $\mathbf{V} = \mathbb{Z}^T\mathbf{H}^T$ a máme

$$\begin{aligned}\hat{\mathbf{P}}^a &= \left(\mathbf{I} - \mathbb{Z}\mathbb{Z}^T\mathbf{H}^T (\mathbf{H}\mathbb{Z}\mathbb{Z}^T\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H} \right) \mathbb{Z}\mathbb{Z}^T \\ &= \mathbb{Z} \left(\mathbf{I} - \mathbb{Z}^T\mathbf{H}^T (\mathbf{H}\mathbb{Z}\mathbb{Z}^T\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbb{Z} \right) \mathbb{Z}^T \\ &= \mathbb{Z} (\mathbf{I} - \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T) \mathbb{Z}^T\end{aligned}$$

Matice uprostřed je typu $N \times N$ a můžeme ji odmocnit a update perturbací provést jako

$$\mathbb{Z} (\mathbf{I} - \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T)^{1/2}$$

Jiná verze square-root filtru (Whitaker and Hamill 2002)

Update průměru (či jiného odhadu skutečného stavu) se provede standardně, update perturbací se řeší použitím jiné matice gainu, $\tilde{\mathbf{K}}$, která je zvolena tak, aby platilo

$$(\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H}) \mathbf{P} (\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H})^T = (\mathbf{I} - \mathbf{K}\mathbf{H}) \mathbf{P}$$

Dá se ukázat, že rovnice má řešení

$$\tilde{\mathbf{K}} = (\mathbf{P}^f \mathbf{H}^T) [(\sqrt{\mathbf{D}})^{-1}]^T (\sqrt{\mathbf{D}} + \sqrt{\mathbf{R}})^{-1} .$$

kde pro \mathbf{V} symetrickou pozitivně def. matici $\sqrt{\mathbf{V}}$ značí matici stejného rozměru takovou, že $\sqrt{\mathbf{V}}(\sqrt{\mathbf{V}})^T = \mathbf{V}$. Update perturbací tedy je

$$\mathbb{Z}^a = (\mathbf{I} - \tilde{\mathbf{K}}\mathbf{H}) \mathbb{Z}.$$

Lokalizace kovariancí

Bud' ρ lokalizační funkce. Update průměru se provede standardně

$$\bar{\mathbf{X}}^a = \bar{\mathbf{X}} + \mathbf{K}(\mathbf{y} - \mathbf{H}\bar{\mathbf{X}}),$$

ale ve vzorci pro Kalman gain se použijí lokalizované matice

$$\mathbf{K}^\circ = \rho \circ (\mathbf{P}^f \mathbf{H}^T) (\mathbf{D}^\circ)^{-1},$$

$$\mathbf{D}^\circ = \rho \circ (\mathbf{H} \mathbf{P}^f \mathbf{H}^T) + \mathbf{R},$$

(symbol \circ značí Schurův součin).

stejně jako ve vzorci pro update perturbací:

$$\tilde{\mathbf{K}}^\circ = \rho \circ (\mathbf{P}^f \mathbf{H}^T) [(\sqrt{\mathbf{D}^\circ})^{-1}]^T (\sqrt{\mathbf{D}^\circ} + \sqrt{\mathbf{R}})^{-1}.$$

Spektrální diagonální aproximace

Idea

Transformovat členy ensamble do spektrálního prostoru ve kterém je odhadovaná kovarianční matice diagonální.

Spektrální diagonální aproximace

Idea

Transformovat členy ensamble do spektrálního prostoru ve kterém je odhadovaná kovarianční matice diagonální.

- transformace:
 - Fourierovy transformace – FFT (DFT), DST, DCT
 - Waveletové transformace – DWT
- použít jenom diagonální prvky výběrové kovarianční matice v spektrálním prostoru

Spektrální diagonální aproximace

Idea

Transformovat členy ensamble do spektrálního prostoru ve kterém je odhadovaná kovarianční matice diagonální.

- transformace:
 - Fourierovy transformace – FFT (DFT), DST, DCT
 - Waveletové transformace – DWT
- použít jenom diagonální prvky výběrové kovarianční matice v spektrálním prostoru
- odhad je pozitivně definitní matice
- potřebný počet členů ensamble by se měl snížit
- výpočetně by to nemělo být horší

Značení

$\mathbf{X}_1, \dots, \mathbf{X}_N$ – členy ensamble (délky n)

$\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_N$ – transformovaný ensabl, $\tilde{\mathbf{X}}_i = \mathbf{T}\mathbf{X}_i$

\mathbf{T} – (ortogonální) transformační matice

$\mathbf{P}, \tilde{\mathbf{P}}$ – skutečné kovarianční matice, $\tilde{\mathbf{P}} = \mathbf{T}\mathbf{P}\mathbf{T}^T$

Značení

$\mathbf{X}_1, \dots, \mathbf{X}_N$ – členy ensamble (délky n)

$\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_N$ – transformovaný ensembl, $\tilde{\mathbf{X}}_i = \mathbf{T}\mathbf{X}_i$

\mathbf{T} – (ortogonální) transformační matice

$\mathbf{P}, \tilde{\mathbf{P}}$ – skutečné kovarianční matice, $\tilde{\mathbf{P}} = \mathbf{T}\mathbf{P}\mathbf{T}^T$

$\mathbf{C}_N, \tilde{\mathbf{C}}_N$ – výběrové kovarianční matice

$\tilde{\mathbf{D}}_N$ – diagonální aproximace v spektrálním prostoru,

$$\tilde{\mathbf{D}}_N = \tilde{\mathbf{C}}_N \circ \mathbf{I} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

Značení

$\mathbf{X}_1, \dots, \mathbf{X}_N$ – členy ensamble (délky n)

$\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_N$ – transformovaný ensembl, $\tilde{\mathbf{X}}_i = \mathbf{T}\mathbf{X}_i$

\mathbf{T} – (ortogonální) transformační matice

$\mathbf{P}, \tilde{\mathbf{P}}$ – skutečné kovarianční matice, $\tilde{\mathbf{P}} = \mathbf{T}\mathbf{P}\mathbf{T}^T$

$\mathbf{C}_N, \tilde{\mathbf{C}}_N$ – výběrové kovarianční matice

$\tilde{\mathbf{D}}_N$ – diagonální aproximace v spektrálním prostoru,

$$\tilde{\mathbf{D}}_N = \tilde{\mathbf{C}}_N \circ \mathbf{I} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

\mathbf{D}_N – spektrální diagonální aproximace v původním prostoru $\mathbf{D}_N = \mathbf{T}^T \tilde{\mathbf{D}}_N \mathbf{T}$

Fourierova transformace

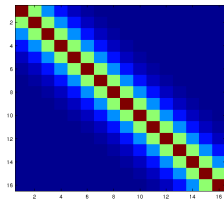
Téměř pravdivé tvrzení

$\tilde{\mathbf{P}}$ je diagonální právě vtedy, když je \mathbf{X} (...) stacionární

Fourierova transformace

Téměř pravdivé tvrzení

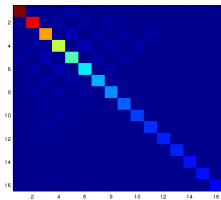
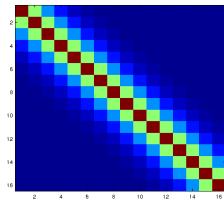
$\tilde{\mathbf{P}}$ je diagonální právě vtedy, když je \mathbf{X} (...) stacionární



Fourierova transformace

Téměř pravdivé tvrzení

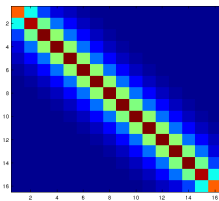
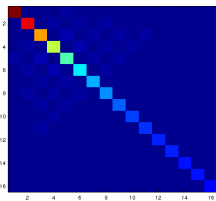
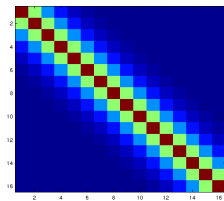
$\tilde{\mathbf{P}}$ je diagonální právě vtedy, když je \mathbf{X} (...) stacionární



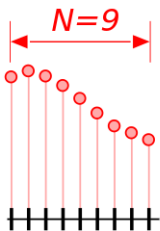
Fourierova transformace

Téměř pravdivé tvrzení

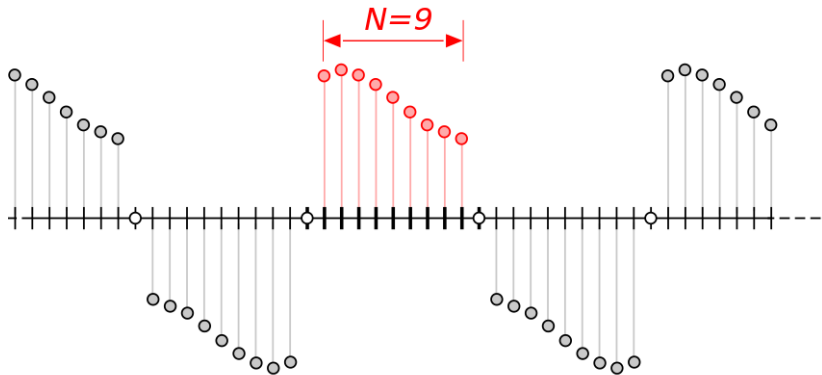
$\tilde{\mathbf{P}}$ je diagonální právě vtedy, když je \mathbf{X} (...) stacionární



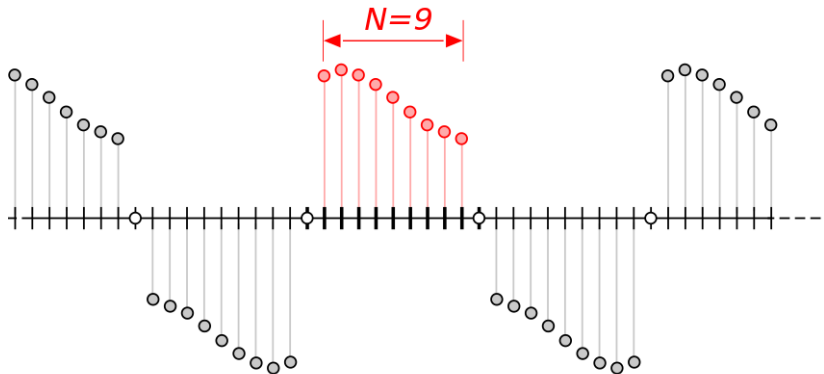
Fourierova transformace



Fourierova transformace



Fourierova transformace



Pravdivé tvrzení

$\tilde{\mathbf{P}}$ je diagonální právě tehdy, když je k -periodické rozšíření \mathbf{X} stacionární

Wavelety

- aproximují libovolnou funkci z L^2

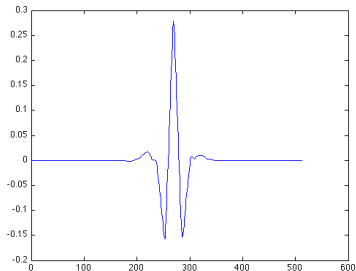
$$V_0 \subset V_1 \subset V_2 \subset V_3 \subset \dots \subset L^2$$

$$V_n = \text{span}(\varphi(2^n x + k)_{k \in \mathbb{Z}})$$

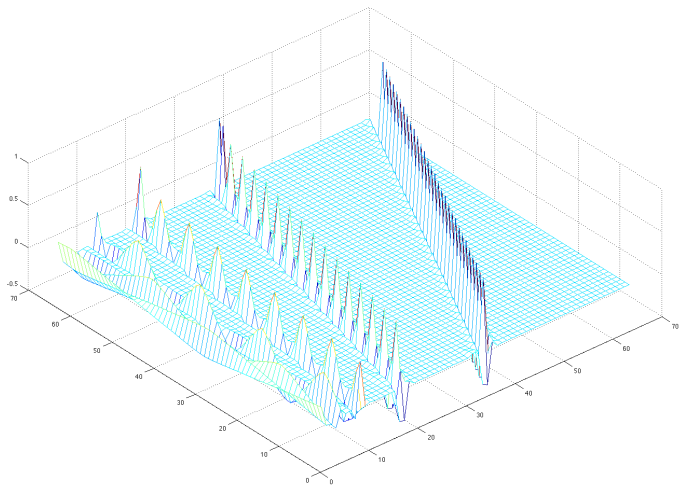
- wavelety jsou (ortogonální) báze prostoru

$$W_n = V_{n+1} / V_n$$

$$W_n = \text{span}(\psi(2^n x + k)_{k \in \mathbb{Z}})$$



Wavelety



Střední chyba odhadu

Jak generovat \mathbf{X} ?

$$\mathbf{X} = \sum_{i=1}^n \lambda_k^{1/2} \theta_k \mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \underbrace{\sum_{i=1}^n \lambda_k \mathbf{u}_k \mathbf{u}_k^T}_{\mathbf{P}})$$

$\theta_k \sim \mathcal{N}(0, 1)$ i.i.d

$\{\mathbf{u}_i\}_{i=1}^n$ spektrální ortonormální báze

$$\tilde{\mathbf{P}} = \mathbf{TPT}^T = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Střední chyba odhadu

$$E\|P - C_N\|_F^2 = E\|P - D_N\|_F^2 + \sum_{i \neq j} \frac{\lambda_i \lambda_j}{N}$$

$$E\|P - D_N\|_F^2 = \frac{2}{N}\|P\|_F^2 = \frac{2}{N} \sum_{i=1}^n \lambda_i^2$$

Střední chyba odhadu

$$E\|P - C_N\|_F^2 = E\|P - D_N\|_F^2 + \sum_{i \neq j} \frac{\lambda_i \lambda_j}{N}$$

$$E\|P - D_N\|_F^2 = \frac{2}{N}\|P\|_F^2 = \frac{2}{N} \sum_{i=1}^n \lambda_i^2$$

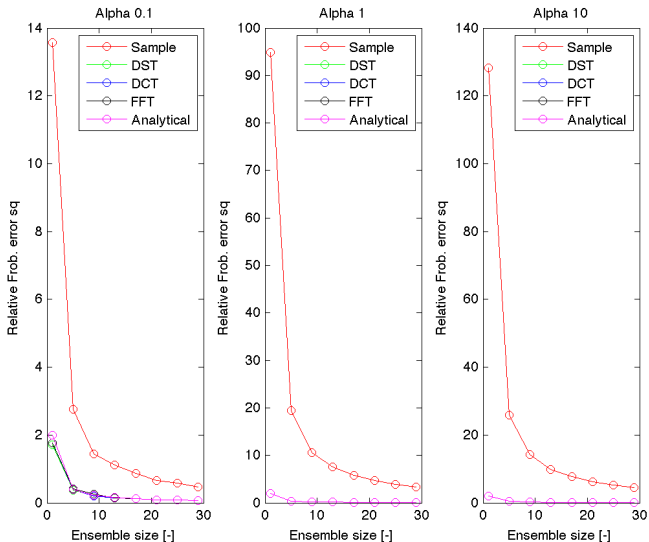
Za předpokladu fixního N a pro $n \rightarrow \infty$:

C_N má konečnou očekávanou chybu když $\{\lambda_i\}_{i=1}^{\infty} \in l^1$

D_N má konečnou očekávanou chybu když $\{\lambda_i\}_{i=1}^{\infty} \in l^2$

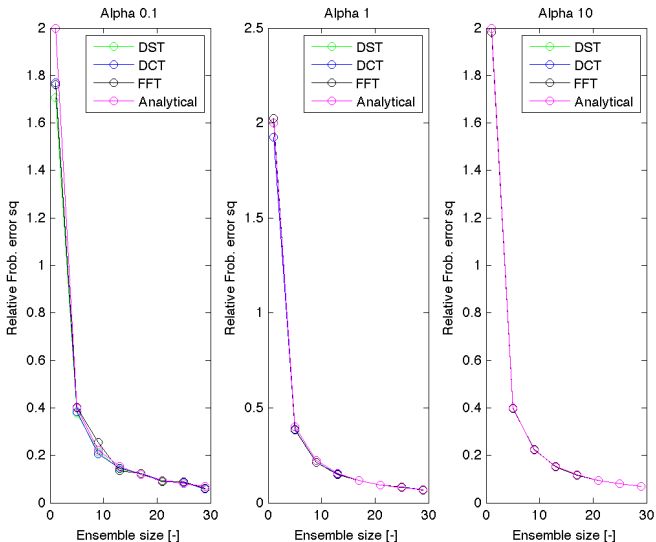
Simulace – průměrná chyba odhadu

Stacionární náhodné pole



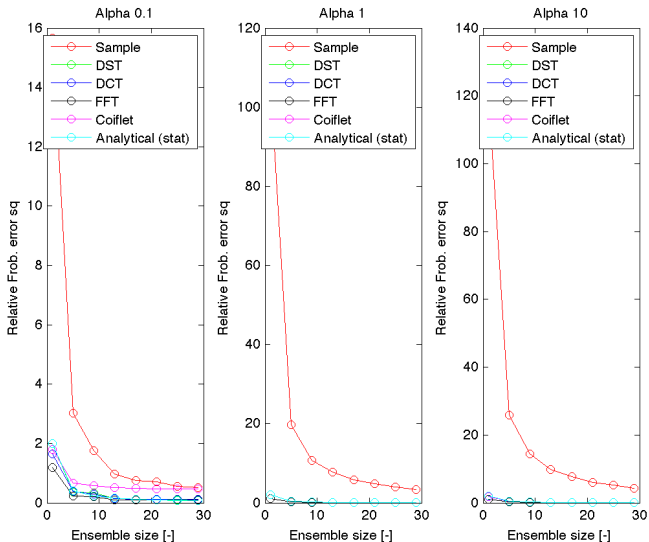
Simulace – průměrná chyba odhadu

Stacionární náhodné pole



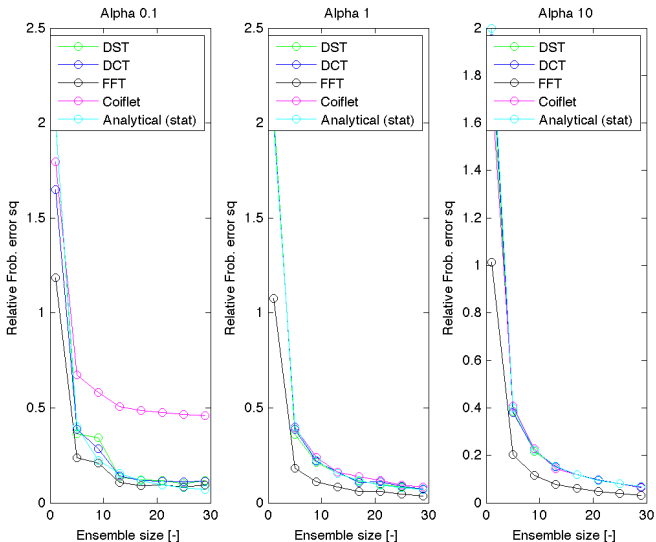
Simulace – průměrná chyba odhadu

Téměř stacionární náhodné pole



Simulace – průměrná chyba odhadu

Téměř stacionární náhodné pole



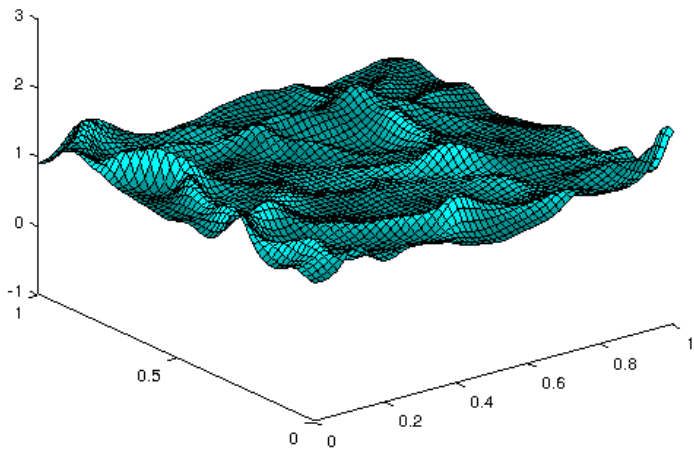
Shallow water equations

$$\frac{\partial \eta}{\partial t} + \frac{\partial(\eta u)}{\partial x} + \frac{\partial(\eta v)}{\partial y} = 0$$

$$\frac{\partial(\eta u)}{\partial t} + \frac{\partial}{\partial x} \left(\eta u^2 + \frac{1}{2} g \eta^2 \right) + \frac{\partial(\eta u v)}{\partial y} = -g \eta \frac{\partial H}{\partial x}$$

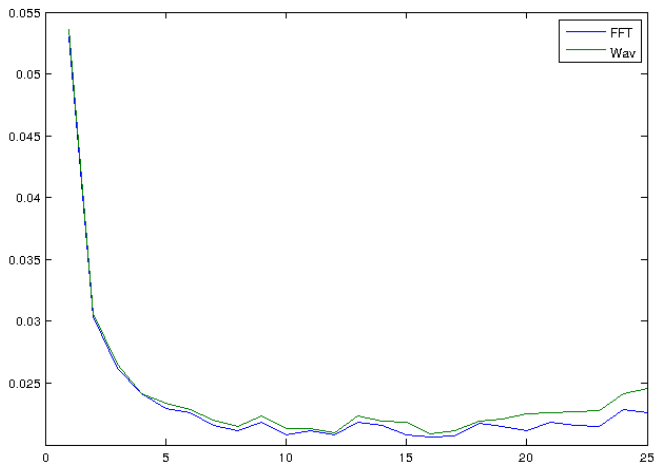
$$\frac{\partial(\eta v)}{\partial t} + \frac{\partial(\eta u v)}{\partial x} + \frac{\partial}{\partial y} \left(\eta v^2 + \frac{1}{2} g \eta^2 \right) = -g \eta \frac{\partial H}{\partial y}.$$

Shallow water equations



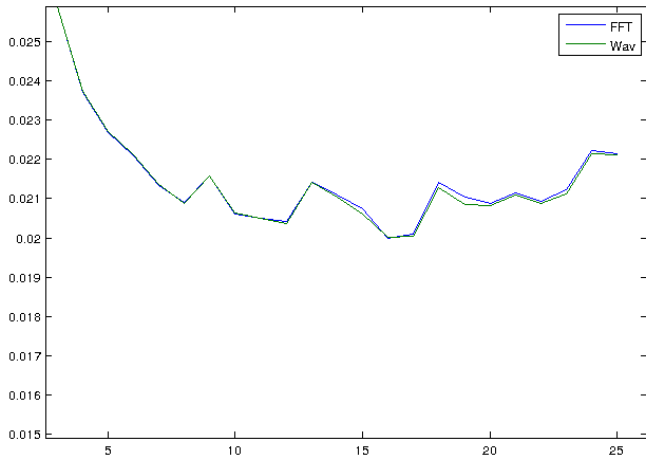
RMSE

Asimilována celá jedna proměnná



RMSE

Asimilována část jedné proměnné



Závěr

- vysoká dimenze stavového vektoru je specifický fenomén pro aplikace v atmosférických vědách
- neustále se objevují nové zdroje dat, které si zaslouží pozornost
- asimilace dat do numerických modelů počasí je hezký zdroj statistických problémů

References

[1] Eben, K., Juruš, P., Resler, J., Belda, M., Pelikán, E., Krüger, B.C. and Keder, J.: An Ensemble Kalman filter for short term forecasting of tropospheric ozone concentrations. *Q. J. R. Met. Soc.* 131 (2005), 3313-3322.

[3] Whitaker, J. S. and Hamill, T. M. (2002): Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.*, 130, 1913–1924