NON-GAUSSIAN DATA ASSIMILATION

STEVEN FLETCHER

COOPERATIVE INSITUTE FOR RESEARCH IN THE ATMOSPHERE,
COLORADO STATE UNIVERSITY

Lecture for the Workshop on Applications of Remotely Sensed Observations in Data Assimilation,
University of Maryland, August 3rd 2007

Colorado State University CIRA

CG/AR DoD Center for Geosciences/Atmospheric Research Colorado State University

Overview of Lecture

1) Descriptive Univariate Statistics

2) Multivariate Distribution Theory

3) Bayes Theorem and Gaussian Data Assimilation

4) Lognormal Distribution, Univariate and Multivariate theory

5) Lognormal Data Assimilation

6) Hybrid Data Assimilation

7) 4D Lognormal Data Assimilation

Who is Steven Fletcher?

B.Sc. (HONS) Mathematics and Statistics, 1998

M.Sc. Numerical Solutions to Differential Equations, 1999 (Prof. Nancy Nichols)

Ph.D. Numerical Weather Prediction, 2004 (Prof. Nancy Nichols, Dr. Ian Roulstone)

All from the Department of Mathematics (Statistics), University of Reading, U. K.

Post-Doc, CIRA/CSU 2004 – 2006, (Dr. Milija Zupanski)
Research Scientist II 2006 - present

Univariate Descriptive Statistics

We shall only be considering the continuous probability distributions for this lecture. The three descriptive statistics for cts. distributions are defined as

The cts. Mean, which is also the first moment of the distribution, it is also the minimum variance estimator and is given by

$$\mu = E(X = x) = \int_a^b x f(x) dx$$

$$x \in [a, b]$$

The next statistic is the median. This is the unbiased estimator. For cts. distributions this is given by

$$x_{med} \text{ is the values such that } \int_{a}^{x_{med}} f(z)dz = \frac{1}{2}$$

The third and final descriptive statistic is the mode. This is the maximum likelihood estimator and is defined by

$$x_{\text{mod}} \text{ is the value of } x \text{ such that } \left.\frac{df}{dx}\right|_{x=x_{\text{mod}}} = 0$$

Univariate normal (Gaussian) distribution

The univariate normal distribution is regularly used to approximate other distributions through the central limit theorem, we do not discuss that here, but is has two definitions. The Standard Normal is defined by

$$f(z)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{z^2}{2}\right\}\ z\in(-\infty,\infty)$$

Whilst the biased version is given by

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}\ x\in(-\infty,\infty)$$

$$\mu=E(x),\quad \sigma^2=E\left(x^2\right)-\left(E(x)^2\right)$$

Multivariate Distribution Theory

Question: How do the three statistic operators transfer to multivariate theory? NOT VERY WELL!!!

MEAN : Intuition says

$$E(\boldsymbol{x}) = \int_a^b \int_c^d \cdots \int_e^f \boldsymbol{x} f(\boldsymbol{x}) d\boldsymbol{x}$$

$$x_1 \in (a,b),\ x_2 \in (c,d), \cdots, x_N \in (e,f)$$

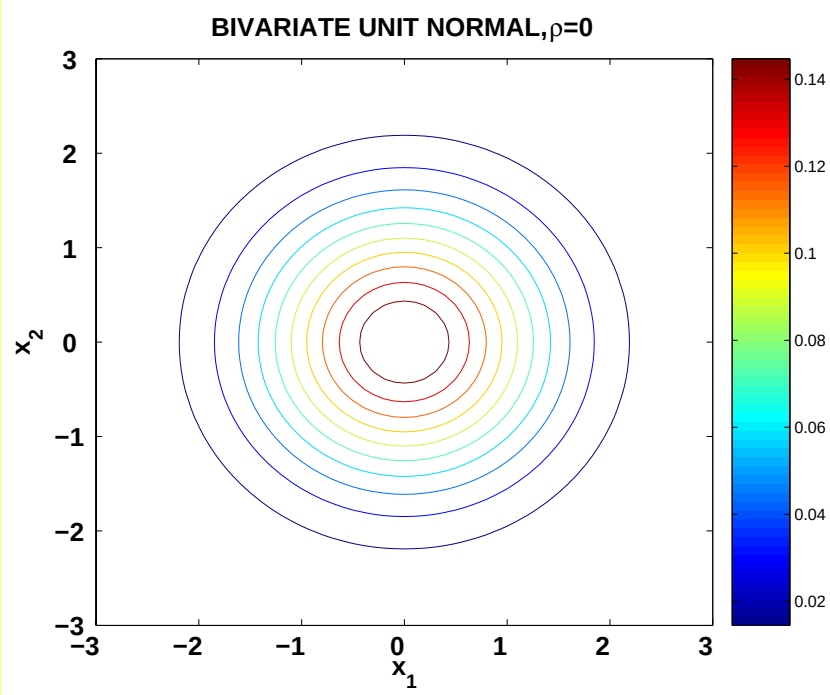
NOT DEFINED IN MATHEMATICS!!!!

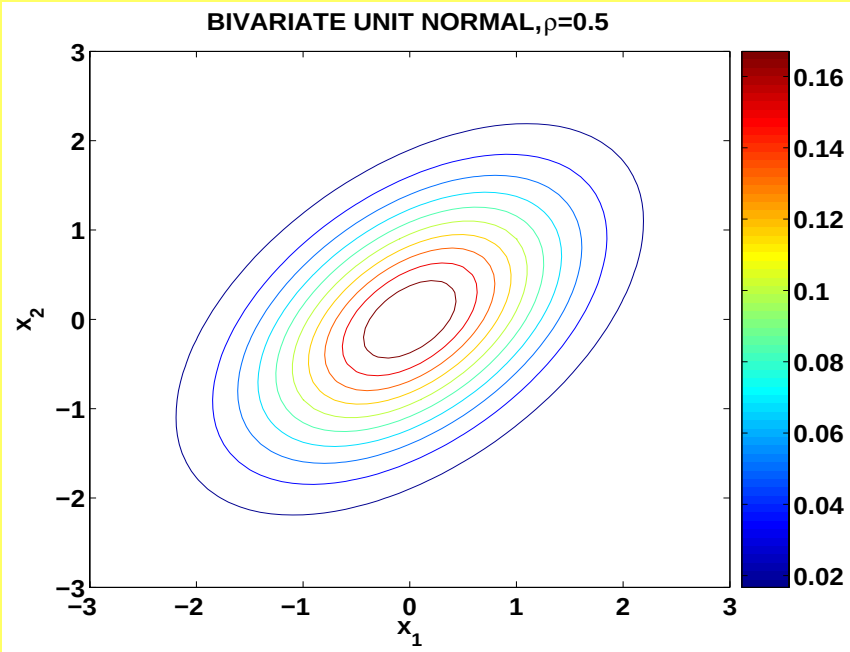
Therefore the expectation of a random vector is the vector of random expectations

MULTIVARIATE MEDIANS

This is the worst of the three statistics. Although it is *unbiased* it is also non-unique, even for the multivariate normal distribution. The multivariate definition is

$$\boldsymbol{x}_{med} = \begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(N) \end{pmatrix} = \int_{a}^{x(1)} \int_{c}^{x(2)} \cdots \int_{e}^{x(N)} f(\boldsymbol{x}) d\boldsymbol{x}$$


BIVARIATE UNIT NORMAL, ρ=0
3
2
1
0
−1
−2
−3
x2
−3
−2
−1
0
1
2
3
x1
0.14
0.12
0.1
0.08
0.06
0.04
0.02


BIVARIATE UNIT NORMAL, ρ=0.5
3
2
1
0
−1
−2
−3
x2
−3
−2
−1
0
1
2
3
x1
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02

Is all hope lost? No

Our saviour is the mode!!!!

There is a simple multivariate extension of the definition of the mode from the univariate case to the multivariate.

$$\boldsymbol{x}_{\text{mod}} \text{ is the value of } \boldsymbol{x} \text{ such that } \left.\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_{\text{mod}}} = \boldsymbol{0}$$

The Multivariate Normal Distribution is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

$$\boldsymbol{x} \in \Re^N$$

BAYES THEOREM AND GAUSSIAN DATA ASSIMILATION

The basis of 3D VAR comes from Lorenc (1986) where he uses Bayes theorem

$$P(A|B) = P(B|A)P(A)$$

with the event A presenting the effect background knowledge and the event B the effects of the observations given the uncertainties with the background states. The right hand side of the equation above becomes the *ANALYSIS DISTRIBUTION* whose mean, mode or median we seek as the *ANALYSIS STATE* OR *TRUE SOLUTION.* For 3D VAR this is written as

$$P(\boldsymbol{x} = \boldsymbol{x}_a | \boldsymbol{y} = \boldsymbol{y}_t) = P(\boldsymbol{y} = \boldsymbol{y}_t | \boldsymbol{x} = \boldsymbol{x}_a) P(\boldsymbol{x} = \boldsymbol{x}_a)$$

$$P(\boldsymbol{x}-\boldsymbol{x}_b) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_b)^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b)\right\}$$

$$P(\boldsymbol{y}=\boldsymbol{y}_t|\boldsymbol{x}=\boldsymbol{x}_a) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x}))\right\}$$

Where we have assumed that the background and observational errors are independent. Therefore multiplying the two distributions above and taking the negative logarithm we obtain the standard 3D VAR cost function

$$J(x) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_b)^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{h}(\boldsymbol{x}))$$

LOGNORMAL DISTRIBUTION: UNIVARIATE AND MULTIVARIATE THEORY

The lognormal distribution is for a class of positive definite random variables, i.e.

$$x \geq 0$$

The pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2} \right\}$$

$$\mu = E(\ln x), \quad \sigma^2 = E\left((\ln x)^2\right) - (E(\ln x))^2$$

PROPERTIES OF THE LOGNORMAL DISTIRBUTION UNIVARIATE

PROPERTY 1: IF $X \sim LN(\mu,\sigma^2)$ THEN ln X IS DISTRIBUTED $N(\mu,\sigma^2)$

PROPERTY 2: THE MODE OF X IS $\exp\{\mu-\sigma^2\}$

PROPERTY 3: THE MEDIAN OF X IS $\exp\{\mu\}$

PROPERTY 4: THE MEAN OF X IS $\exp\left\{\mu+\frac{\sigma^2}{2}\right\}$

PROPERTY 5: THE LOGNORMAL DISTRIBUTION CAN NOT BE UNIQUELY DETERMINED BY ITS MOMENTS!!!

Colorado State University CIRA

NATIONAL SCIENCE FOUNDATION NSF

MULTIVARIATE LOGNORMAL DISTRIBUTION

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left(\prod_{i=1}^{N} \left(\frac{1}{x_i} \right) \right) \times$$

$$\exp\left\{ -\frac{1}{2} (\ln \boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\ln \boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

$$x_{mea}(i) = E(x_i) = \exp\left\{ \mu_i + \frac{\sigma^2}{2} \right\}, i = 1,2,\ldots,N$$

$$\boldsymbol{x}_{med} = \exp\{\boldsymbol{\mu}\},$$

$$\boldsymbol{x}_{\text{mod}} = \exp\left\{\boldsymbol{\mu} - \boldsymbol{\Sigma}^T \boldsymbol{1}\right\}, \boldsymbol{1}^T = (1 \quad 1 \quad \cdots \quad 1)$$

PROPERTIES OF THE LOGNORMAL DISTRIBUTION

MULTIVARIATE

PROPERTY 1: MEDIAN IS NON-UNIQUE

PROPERTY 2: MOMENTS DO NOT DETERMINE THE DISTRIBUTION UNIQUELY

PROPERTY 3: MEAN IS INDEPENDENT OF COVARIANCE AND IS UNDOUNDED WITH RESPECT TO THE VARIANCE

PROPERTY 4: MODE IS BOUNDED AND FINITE WITH RESPECT TO THE VARIANCE.

PROPERTY 5: MODE IS UNIQUE!!!

Colorado State University CIRA

NATIONAL SCIENCE FOUNDATION NSF

LOGNORMAL DATA ASSIMILATION

FLETCHER AND ZUPANSKI (2006a)

BY FOLLOWING LORENC (1986) AND COHN (1997) WE CAN DEFINE A COST FUNCTION FOR LOGNORMAL OBSERVATION ERRORS.

$$J(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x_b}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{x}-\boldsymbol{x_b}\right) + \frac{1}{2}\left(\ln \boldsymbol{y} - \ln \boldsymbol{h}(\boldsymbol{x})\right)^T R^{-1}\left(\ln \boldsymbol{y} - \ln \boldsymbol{h}(\boldsymbol{x})\right) + \sum_{i=1}^{N_o}\left(\ln y_i - \ln h_i(\boldsymbol{x})\right) \qquad (1)$$

WHERE THE ERRORS ARE DEFINED BY

$$\varepsilon_b = \boldsymbol{x}-\boldsymbol{x_b} \propto N(\boldsymbol{0},\boldsymbol{B}), \quad AND \quad \varepsilon_o = \frac{\boldsymbol{y}}{\boldsymbol{h}(\boldsymbol{x})} \propto LN(\boldsymbol{0},\boldsymbol{R}) \qquad (2)$$

Colorado State University CIRA

NATIONAL SCIENCE FOUNDATION NSF

MOTIVATION

UNDERLYING ASSUMPTION IN VARIATIONAL, KALMAN FILTER AND ENSEMBLE DATA ASSIMILATION IS THAT THE VARIABLES, OBSERVATIONS AND HENCE ERRORS ARE NORMAL (GAUSSIAN) DISTRIBUTED, LORENC (1986), KALMAN (1960), EVENSEN (1994).

QUESTION 1: IS THIS TRUE FOR ALL SCALES IN THE ATMOSPHERE?

QUESTION 2: IS THIS TRUE FOR ALL TYPES OF OBSERVATIONS?

Question 1

Synoptic scale: Are all the synoptic variables Gaussian? NO

Humidity: To assimilate this variable we have to use the logarithm of the variable (Polavarapu *et al* 2005). This indicates that this variable is LOGNORMAL!!

Wind component: Combined with the moisture flux then this variable is showing sign that the wind components may be LOGNORMAL!! (Raymond 1997)

Comment: These variables are both positive definite!!

Question 1

Meso-scale: Are all meso-scale variables Gaussian? NO

In the paper by Miles *et al.* (2000) there is a large summary of cloud variables which are not Gaussian, specifically LOGNORMAL and GAMMA. As early as the 1970s rain and cloud variables had been identified as LOGNORMAL, Mielke *et al. (*1977)

Question 2

Are all observations Gaussian? NO

Direct Observations: The variables which we have already mentioned are not Gaussian and therefore a direct observation of them is also not.

Retrievals: By having to take the logarithm of the humidity to apply 1D Variational data assimilation (1D VAR) then the observation is not Gaussian. It is LOGNORMAL!!

Radiances: By having to *bias correct* the solution for the VAR scheme to work implies Gaussian assumption has been violated, (Derber and Wu, (1998), Harris and Kelly (2001)).

Question 2

Optical Depth: In Stephens *et al.* 2002 the pdf of this variable is presented, clearly showing a **LOGNORMAL** structure.

Infra-Red Flux Differences: From the same paper.

Cloud base height: In Sengupta *et al.* 2004 these observations shows signs of a lognormal structure.

Liquid water path: Same paper, shows a sharp positive skewness associated with a **LOGNORMAL** distribution with large variance.

Brightness Temperature: Work currently being done with the GOES-R sounder (Grosso and Sengupta) shows, when scaled to the main region of observation, then this is **LOGNORMAL**.

CURRENT TECHNIQUES

CURRENTLY THERE ARE TWO MECHANISM WHICH ARE USED TO OVERCOME THIS PROBLEM.

TECHNIQUE 1: TRANSFOMATION. IF $X \sim LN(\mu,\sigma^2)$ THEN $\ln X \sim N(\mu,\sigma^2)$. NOTE: THE MEAN AND VARIANCE ARE UNCHANGED.

TECHNIQUE 2: FORCE THE GAUSSIAN ASSUMPTION AND BIAS CORRECT.

A THIRD TECHNIQUE WHICH IS EMPLOYED IN OTHER FIELDS IS TO USE A GAUSSIAN SUM FILTER.

A FOURTH APPROACH IS TO USE A MAXIMUM ENTROPHY CONDITION.

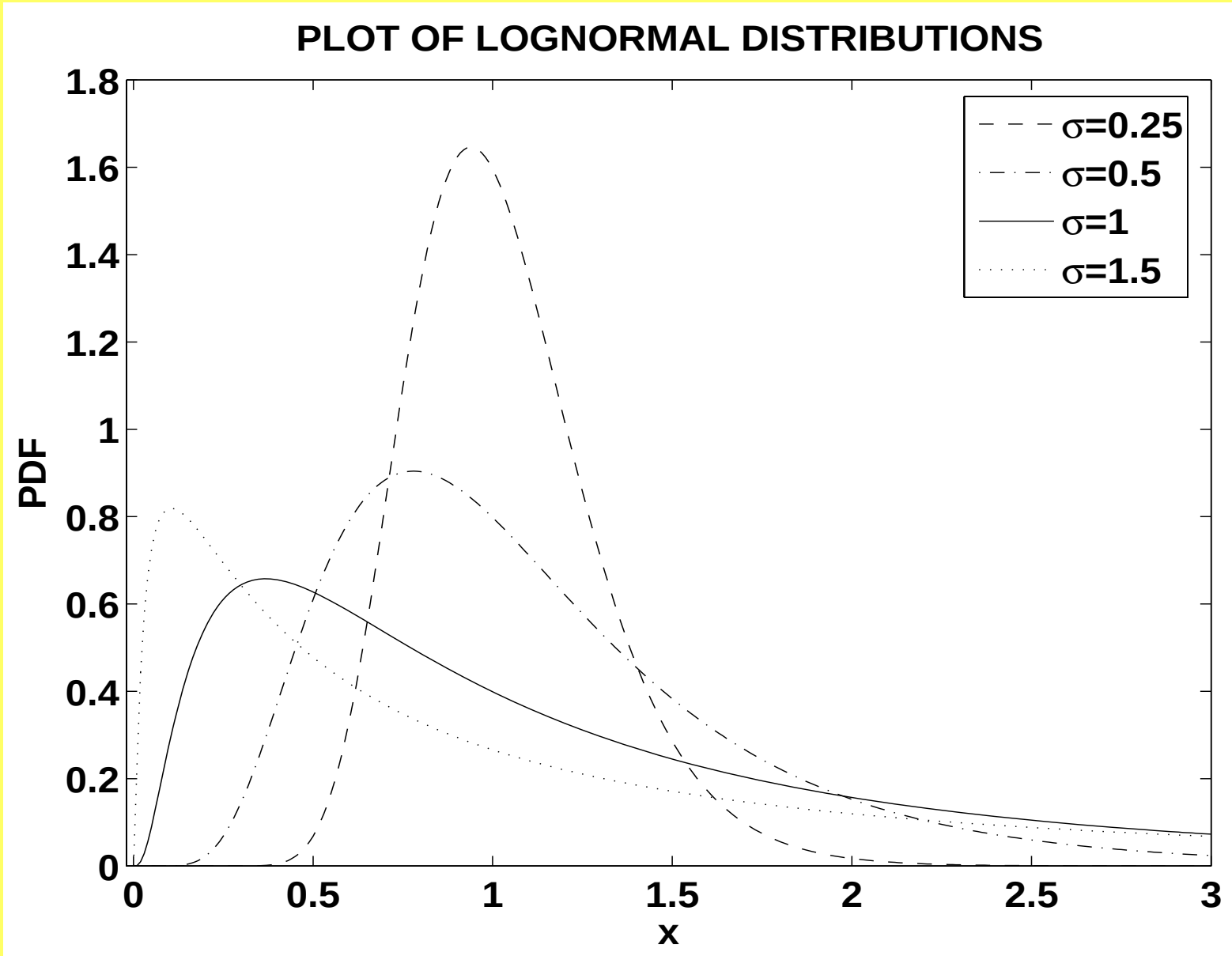
PROBLEMS ASSOCIATED WITH CURRENT TECHNIQUES

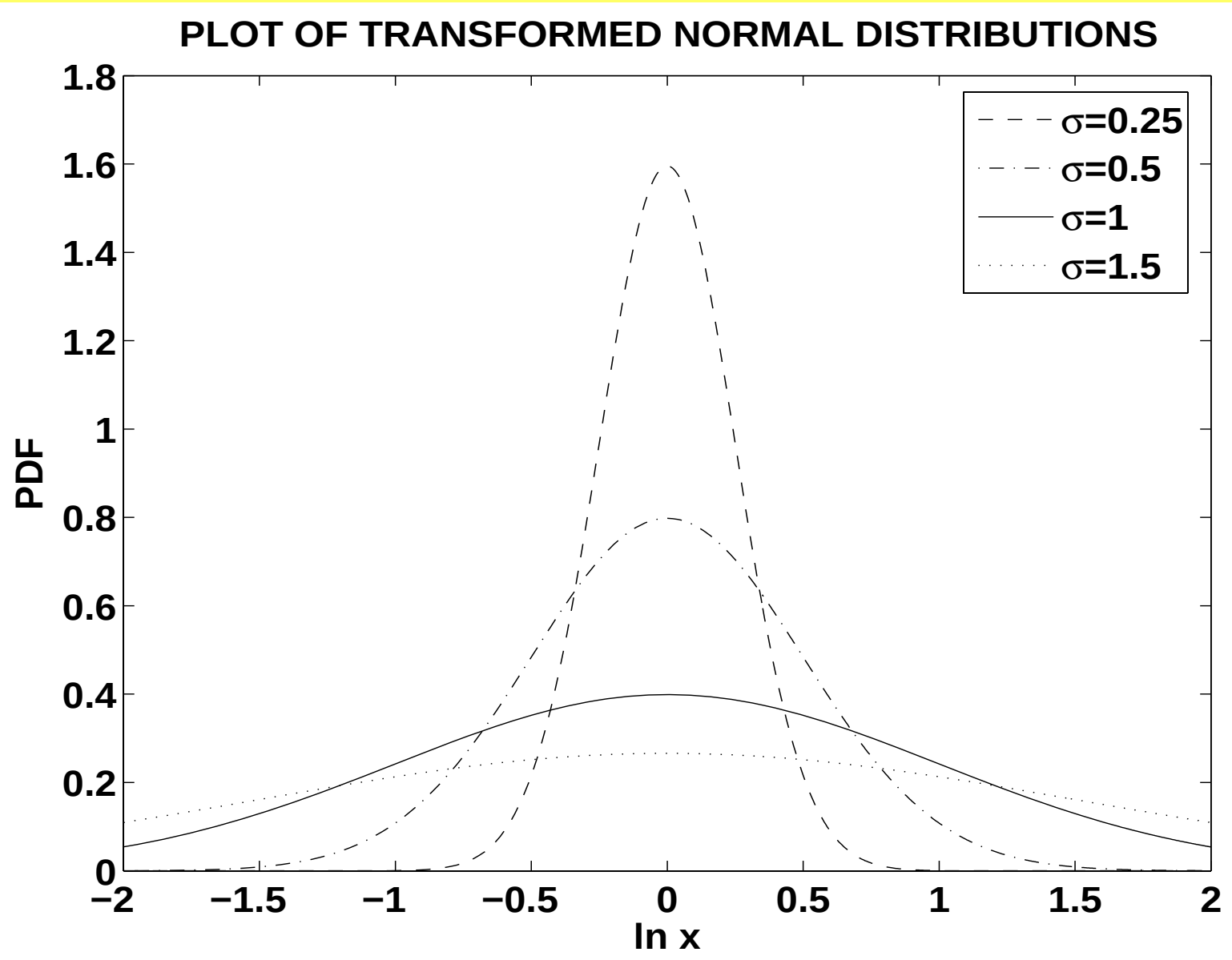
TRANSFORMATION:

IMPACT 1: THE STATISTIC WHICH IS FOUND IS THE MEDIAN NOT THE MODE NOR THE MEAN.

IMPACT 2: WHEN TRANSFORMING ALL HIGHER ORDER MOMENT IMFORMATION IS LOST.

IMPACT 3: THE ANALYSIS STATE FOUND WILL OVER ESTIMATE THE TRUE 'MOST LIKELY STATE'


PLOT OF LOGNORMAL DISTRIBUTIONS
PDF
x
σ=0.25
σ=0.5
σ=1
σ=1.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0
0.5
1
1.5
2
2.5
3


PLOT OF TRANSFORMED NORMAL DISTRIBUTIONS
PDF
ln x
σ=0.25
σ=0.5
σ=1
σ=1.5
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2

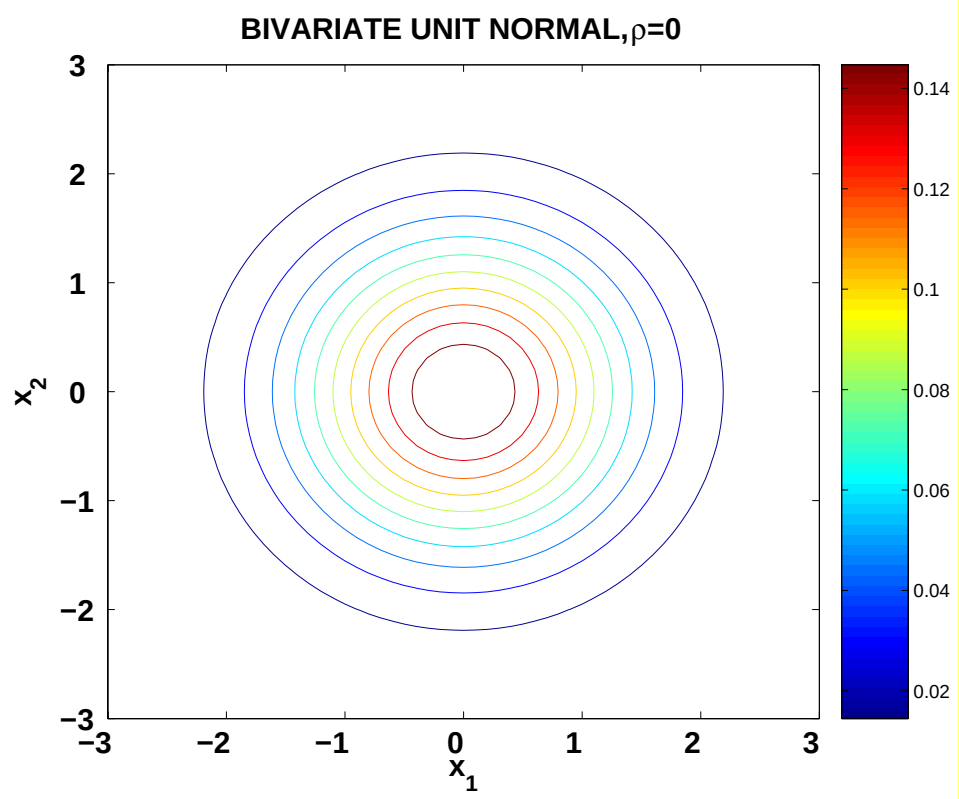
PROBLEMS ASSOCIATED WITH CURRENT TECHNIQUES

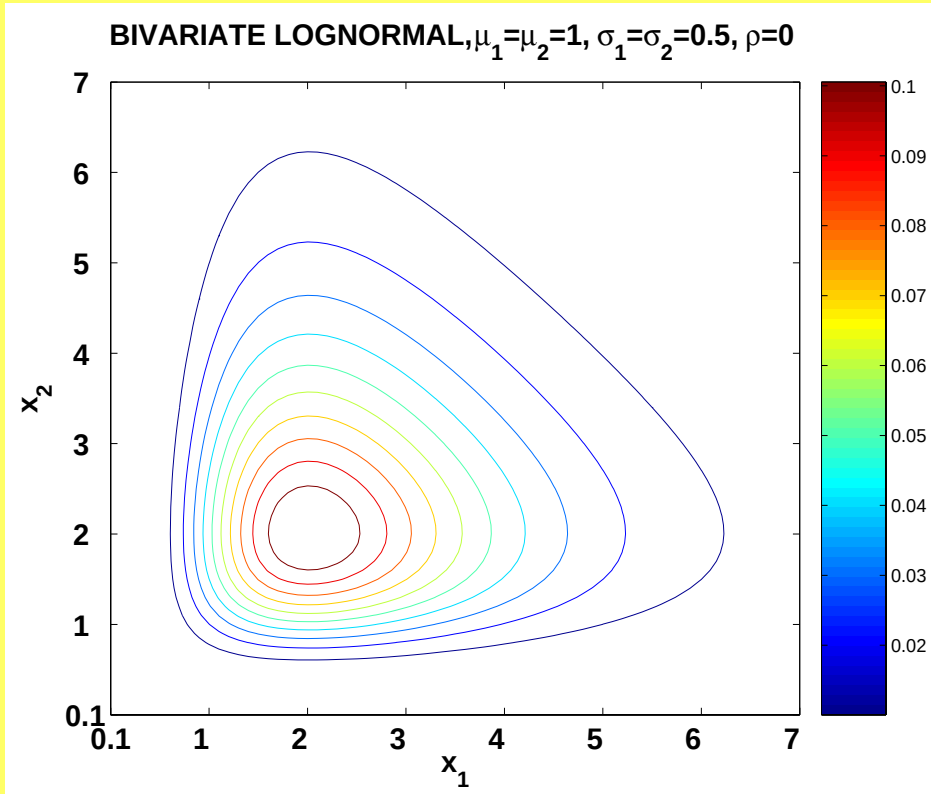
FORCED GAUSSIAN:

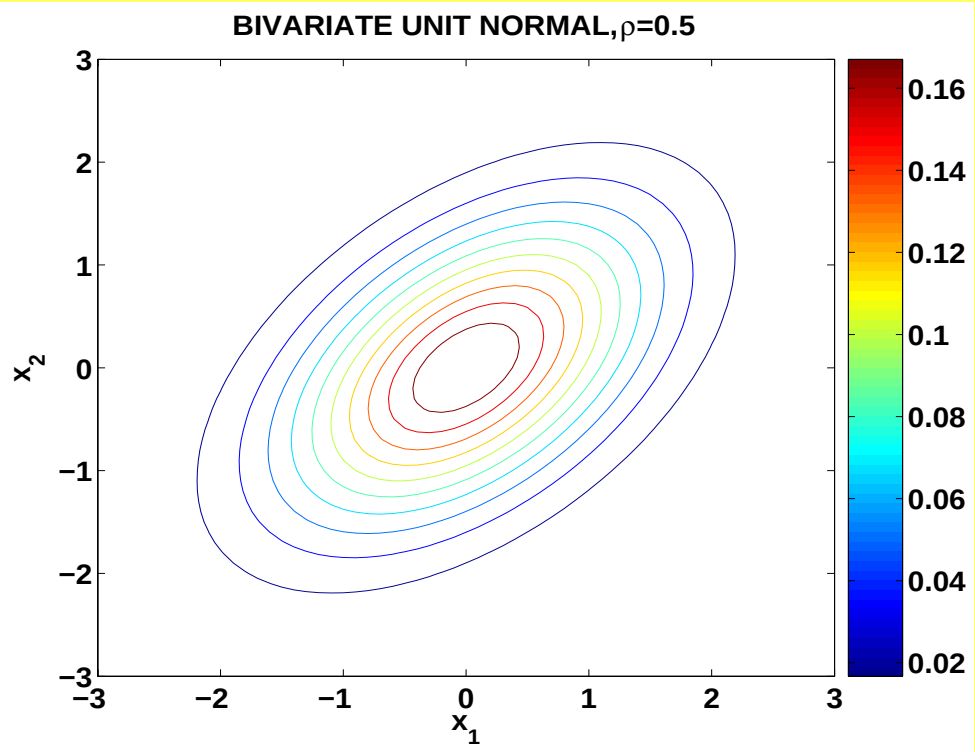
IMPACT 1: WRONG PROBABILITIES ASSIGNED TO THE OUTLIERS.

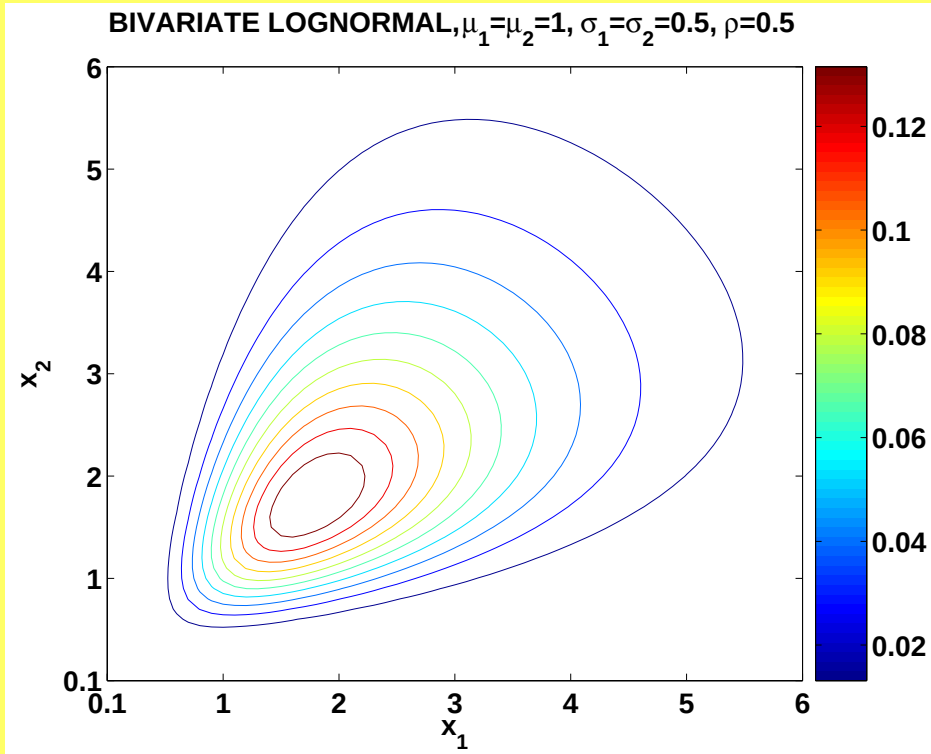
IMPACT 2: PROBABILITIES ASSIGNED TO UNPHYSICAL VALUES.

IMPACT 3: WRONG STATISTICS USED TO APPROXIMATE THE VARIABLE'S DISTRIBUTION.


BIVARIATE UNIT NORMAL, ρ=0
3
2
1
0
−1
−2
−3
x2
−3
−2
−1
0
1
2
3
x1
0.14
0.12
0.1
0.08
0.06
0.04
0.02


BIVARIATE LOGNORMAL, μ1=μ2=1, σ1=σ2=0.5, ρ=0
7
6
5
4
3
2
1
0.1
x2
0.1
1
2
3
4
5
6
7
x1
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02


BIVARIATE UNIT NORMAL, ρ=0.5
3
2
1
0
−1
−2
−3
x2
−3
−2
−1
0
1
2
3
x1
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02


BIVARIATE LOGNORMAL, μ1=μ2=1, σ1=σ2=0.5, ρ=0.5
6
5
4
3
2
1
0.1
x2
0.1
1
2
3
4
5
6
x1
0.12
0.1
0.08
0.06
0.04
0.02

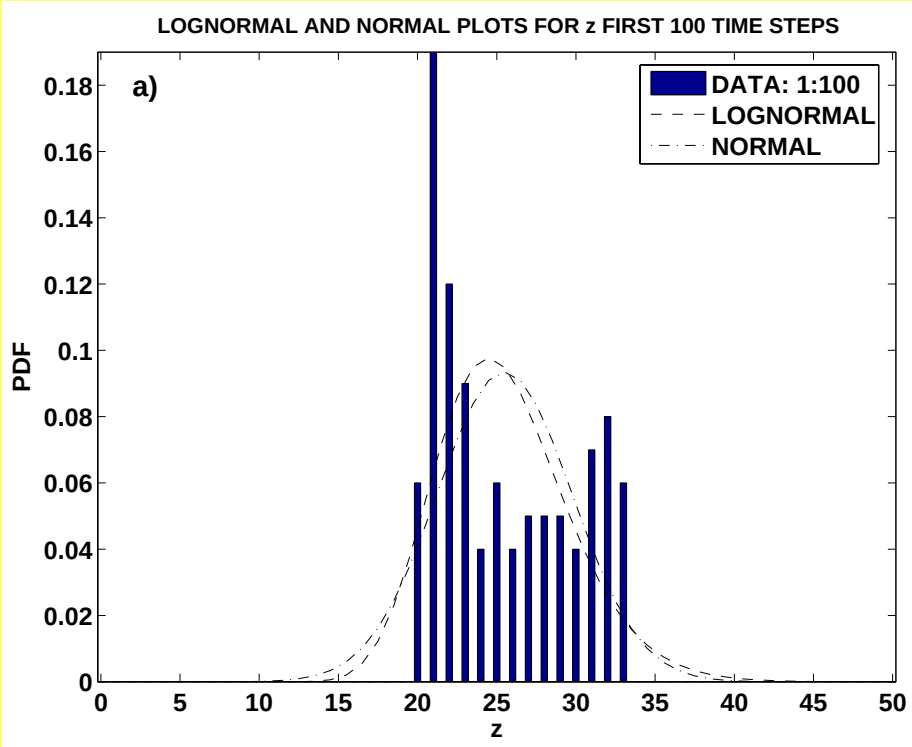
EXAMPLE WITH THE LORENZ'63 MODEL

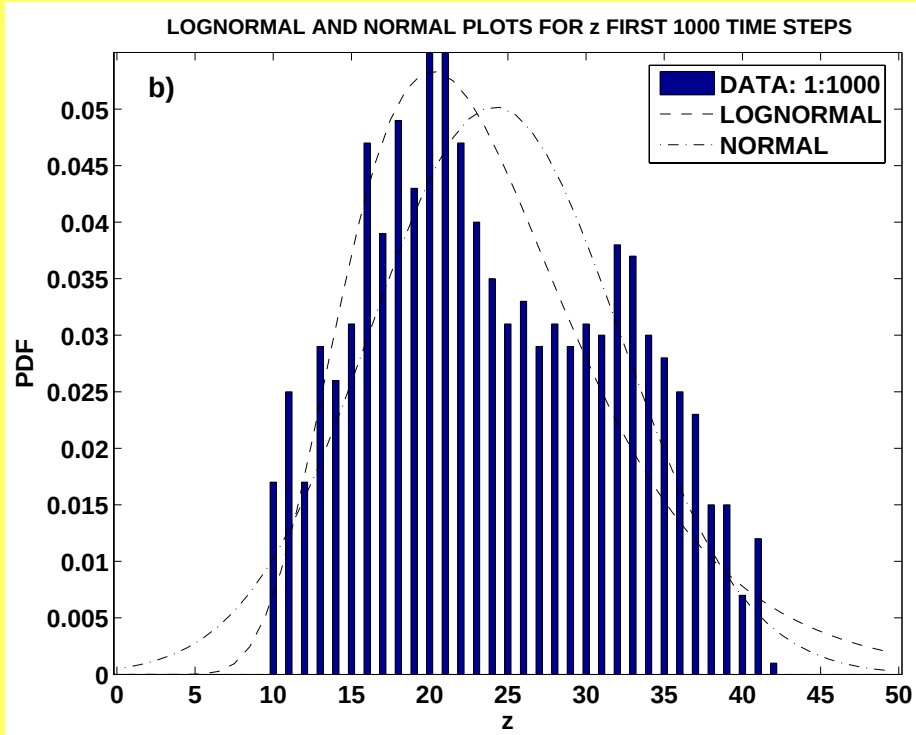
THE MODEL CONSIST OF THE COUPLED SYSTEM OF THREE NON-LINEAR PDES

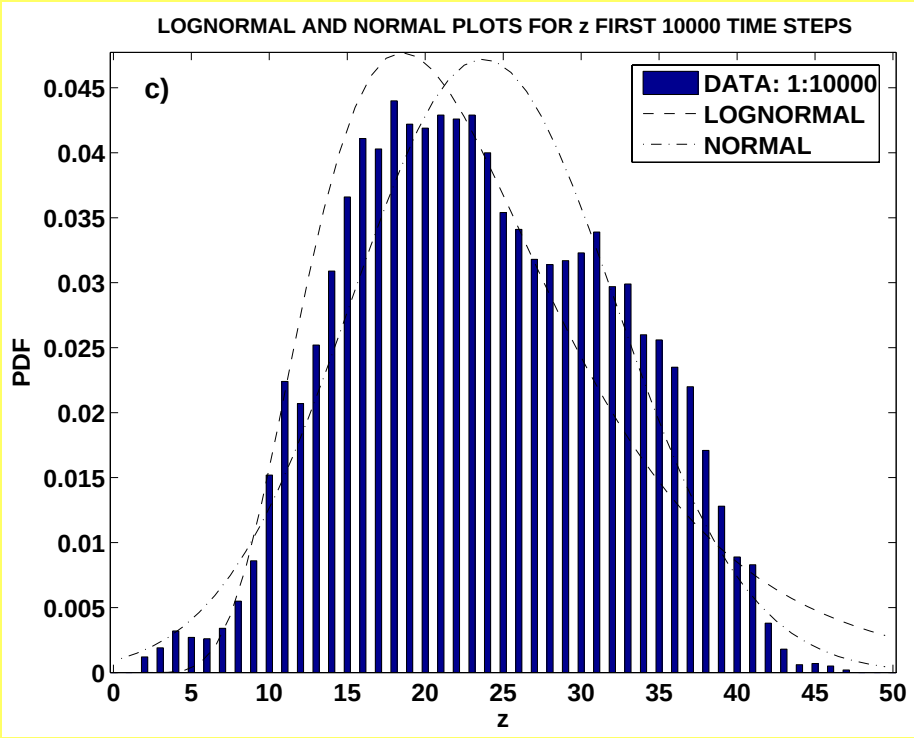
$$\dot{x} = -\sigma x + \sigma y$$
$$\dot{y} = -xz + \rho x - y$$
$$\dot{z} = xy - \beta z$$

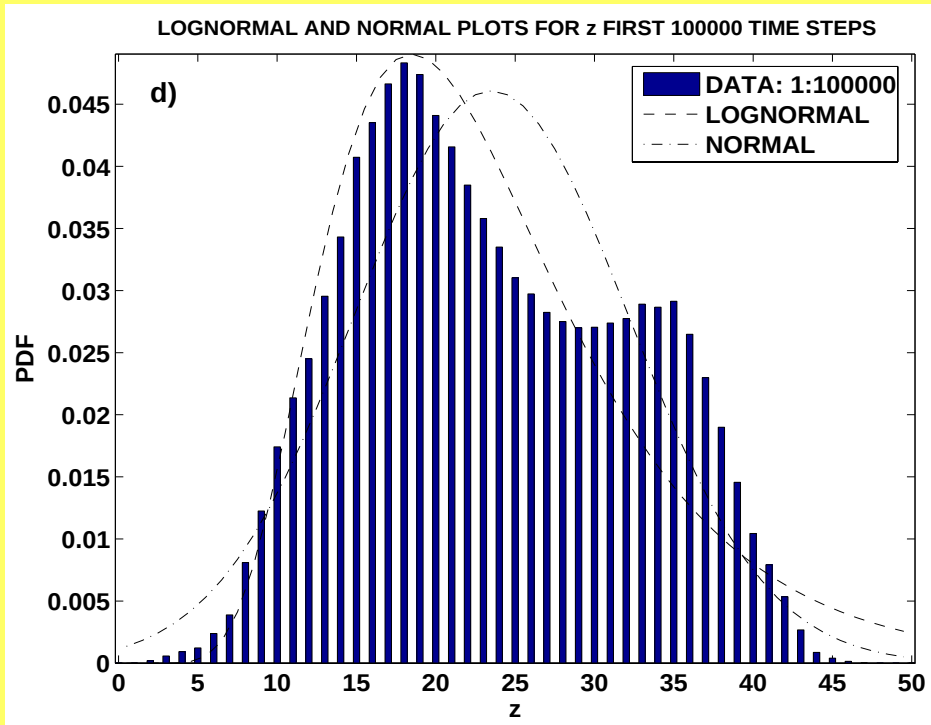
$$\beta = \frac{8}{3}, \quad \sigma = 10 \quad \text{AND} \quad \rho = 28$$

$$x_0 = -5.4458, \quad y_0 = -5.4841 \quad \text{AND} \quad z_0 = 22.5606$$


LOGNORMAL AND NORMAL PLOTS FOR z FIRST 100 TIME STEPS
a)
DATA: 1:100
LOGNORMAL
NORMAL
PDF
z
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0
5
10
15
20
25
30
35
40
45
50


LOGNORMAL AND NORMAL PLOTS FOR z FIRST 1000 TIME STEPS
b)
DATA: 1:1000
LOGNORMAL
NORMAL
PDF
z
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05
0
5
10
15
20
25
30
35
40
45
50


LOGNORMAL AND NORMAL PLOTS FOR z FIRST 10000 TIME STEPS
c)
DATA: 1:10000
LOGNORMAL
NORMAL
PDF
z
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0
5
10
15
20
25
30
35
40
45
50


LOGNORMAL AND NORMAL PLOTS FOR z FIRST 100000 TIME STEPS
d)
DATA: 1:100000
LOGNORMAL
NORMAL
PDF
z
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0
5
10
15
20
25
30
35
40
45
50

LOGNORMAL DATA ASSIMILATION

We start by consider which statistic to use to best represent the underlying analysis pdf.

The three descriptive statistics are the mode 'most likely state', median 'unbiased state' and the mean 'minimum variance'.

Unlike with the Gaussian distribution and other symmetric distributions these three statistics are not identical so which one to use?

HYBRID DISTRIBUTION

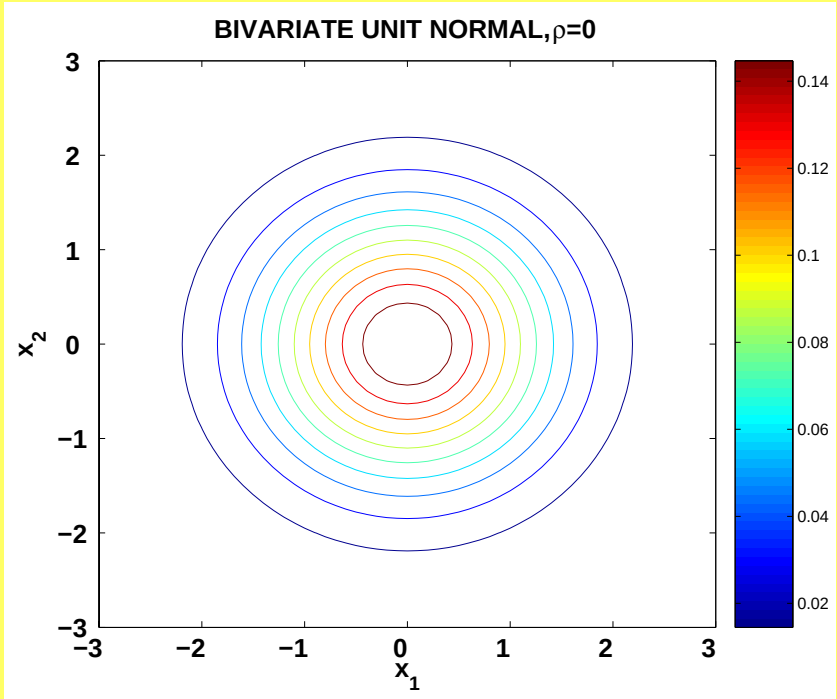
FLETCHER AND ZUPANSKI (2006b)

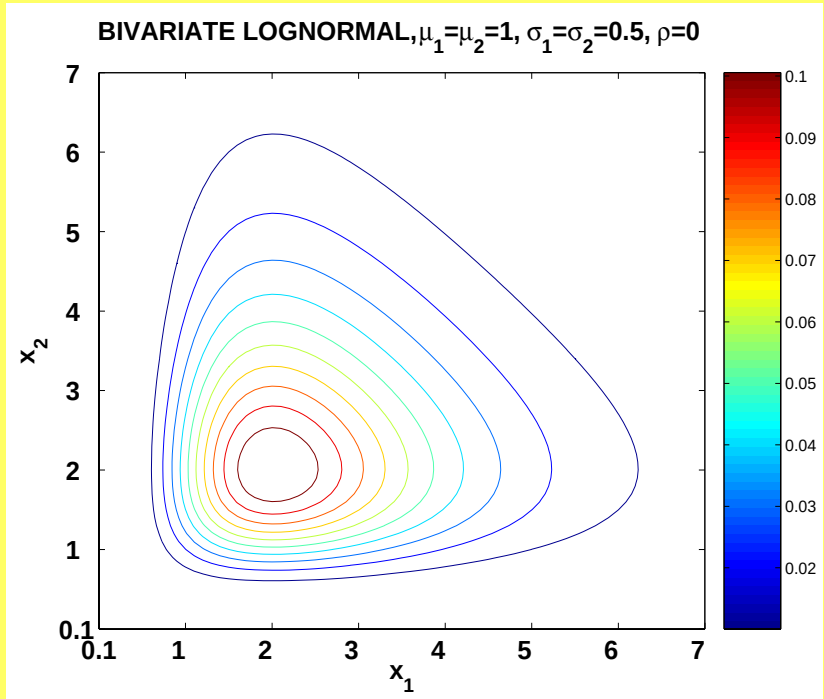
CAN DEFINE A HYBRID NORMAL-LOGNORMAL MULTIVARIATE PROBABILITY DENSITY FUNCTION OF THE FORM

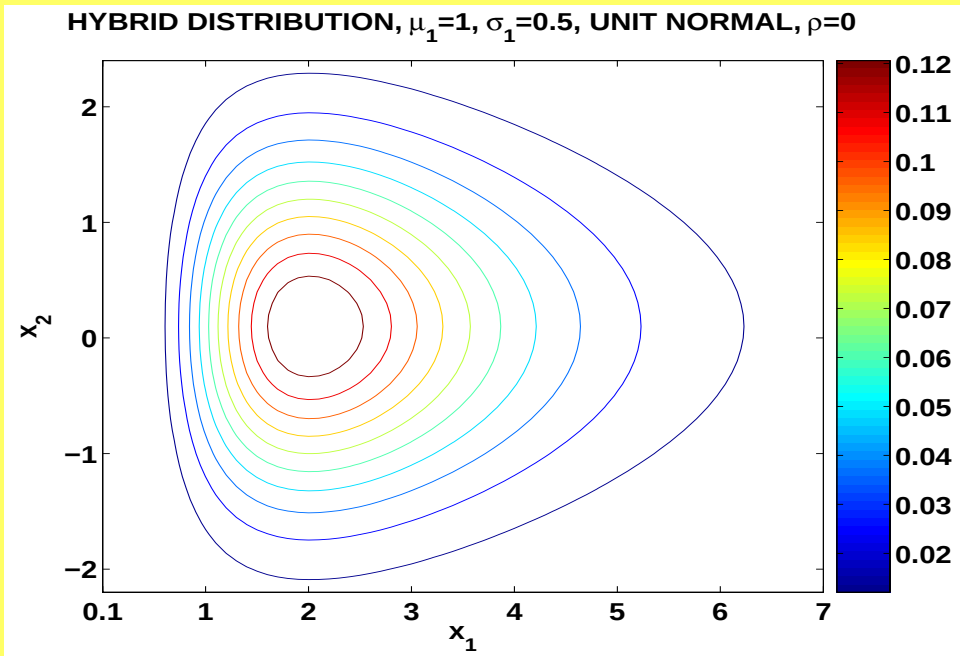
$$f_{p,q}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{R}|^{\frac{1}{2}}} \left(\prod_{i=p+1}^{N} \frac{1}{x_i} \right) \exp\left\{ -\frac{1}{2} (\hat{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{R}^{-1} (\hat{\boldsymbol{x}} - \boldsymbol{\mu}) \right\} \quad (3)$$

WHERE

$$\hat{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x}_p \\ \ln \boldsymbol{x}_q \end{pmatrix}$$


BIVARIATE UNIT NORMAL, ρ=0
x1
x2


BIVARIATE LOGNORMAL, μ1=μ2=1, σ1=σ2=0.5, ρ=0
x1
x2


HYBRID DISTRIBUTION, μ1=1, σ1=0.5, UNIT NORMAL, ρ=0
x1
x2

HYBRID ASSIMILATION

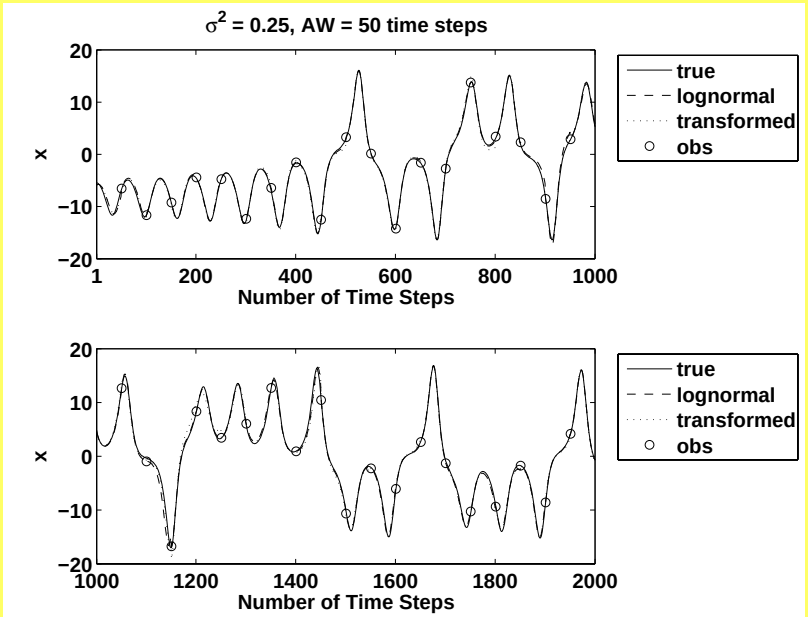
FLETCHER AND ZUPANSKI 2006b

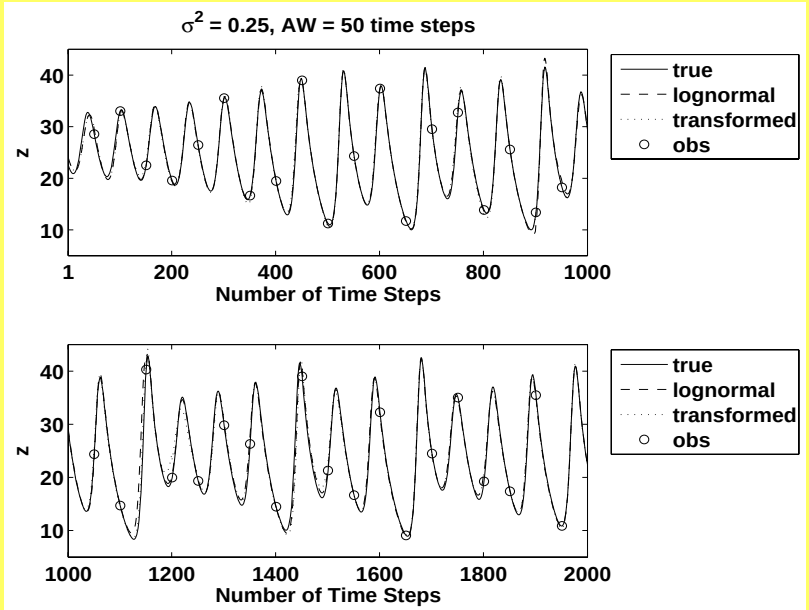
FROM THE DISTRIBUTION DEFINED IN (3) IT IS POSSIBLE TO DEFINE A COST FUNCTION FOLLOWING THE DERIVATION IN LORENC (1986). THIS IS DEFINED WITH A HYBRID BACKGROUND AS

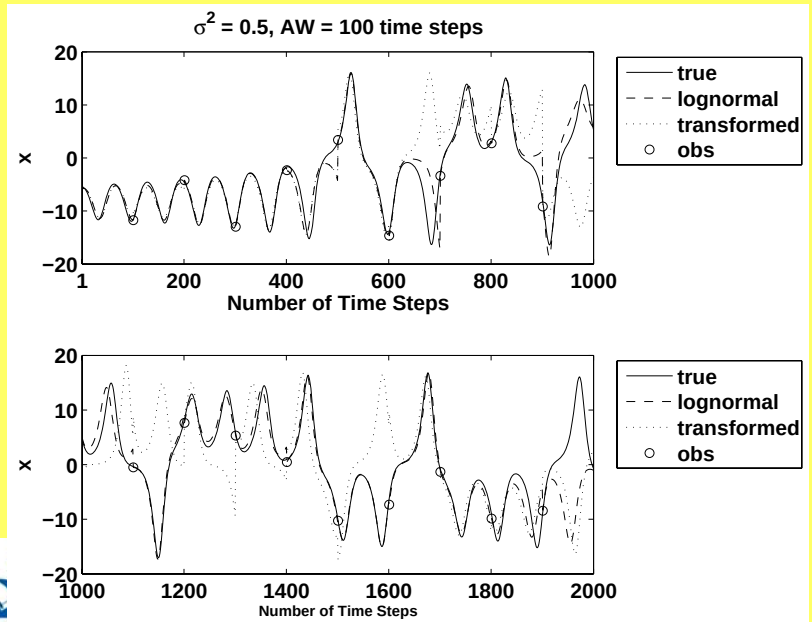
$$J(\boldsymbol{x}) = \frac{1}{2}\hat{\boldsymbol{\varepsilon}}_b^T \boldsymbol{B}^{-1}\hat{\boldsymbol{\varepsilon}}_b + \frac{1}{2}\hat{\boldsymbol{\varepsilon}}_o^T \boldsymbol{R}^{-1}\hat{\boldsymbol{\varepsilon}}_o + \sum_{i=p_1+1}^{N}\hat{\varepsilon}_{bi} + \sum_{j=p_2+1}^{N_o}\hat{\varepsilon}_{oj} \qquad (4)$$

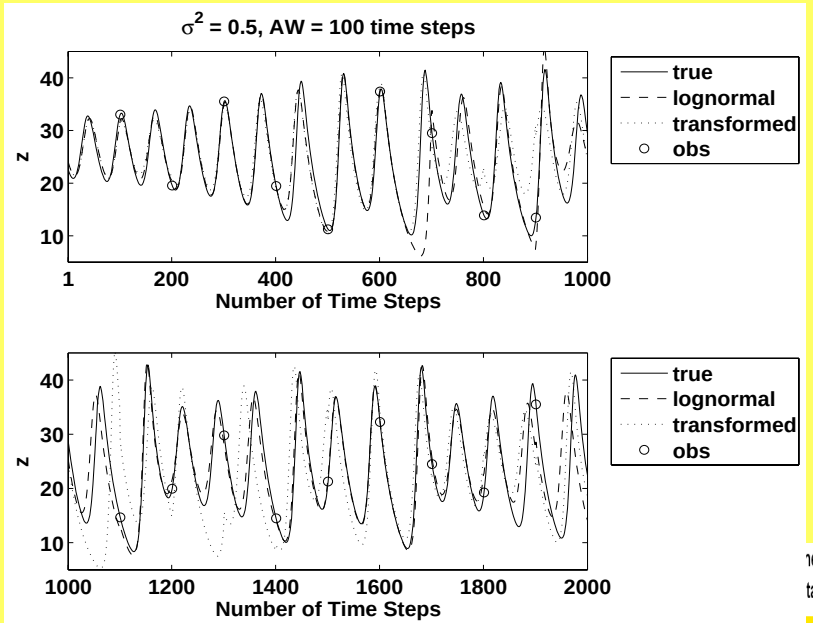
WHERE

$$\hat{\boldsymbol{\varepsilon}}_b = \begin{pmatrix} \boldsymbol{x}_{p_1} - \boldsymbol{x}_{b,p_1} \\ \ln \boldsymbol{x}_{q_1} - \ln \boldsymbol{x}_{b,q_1} \end{pmatrix} \quad \hat{\boldsymbol{\varepsilon}}_o = \begin{pmatrix} \boldsymbol{y}_{p_2} - \boldsymbol{h}_{p_2}(\boldsymbol{x}) \\ \ln \boldsymbol{y}_{q_2} - \ln \boldsymbol{h}_{q_2}(\boldsymbol{x}) \end{pmatrix} \qquad (5)$$



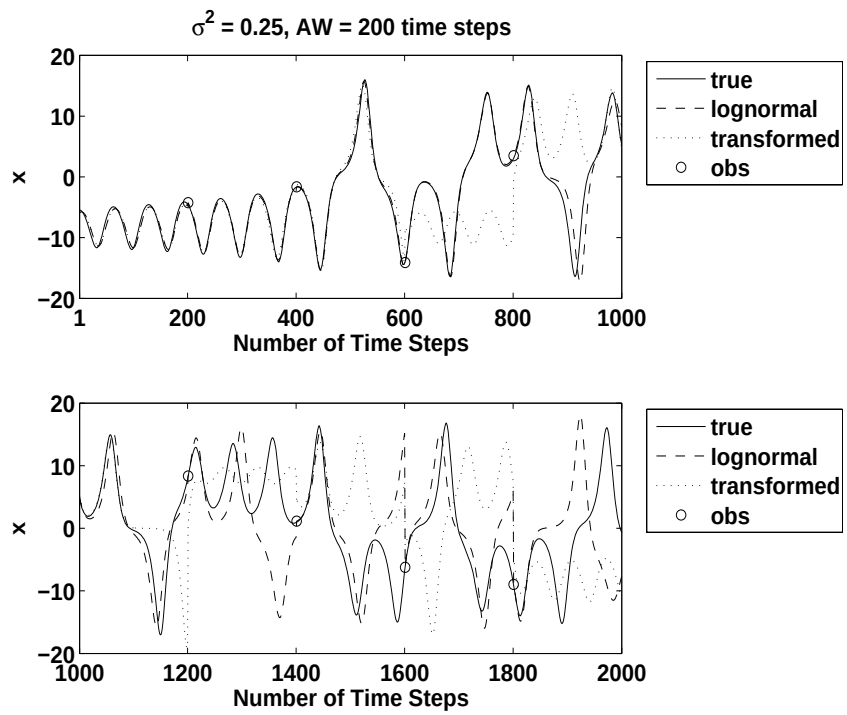


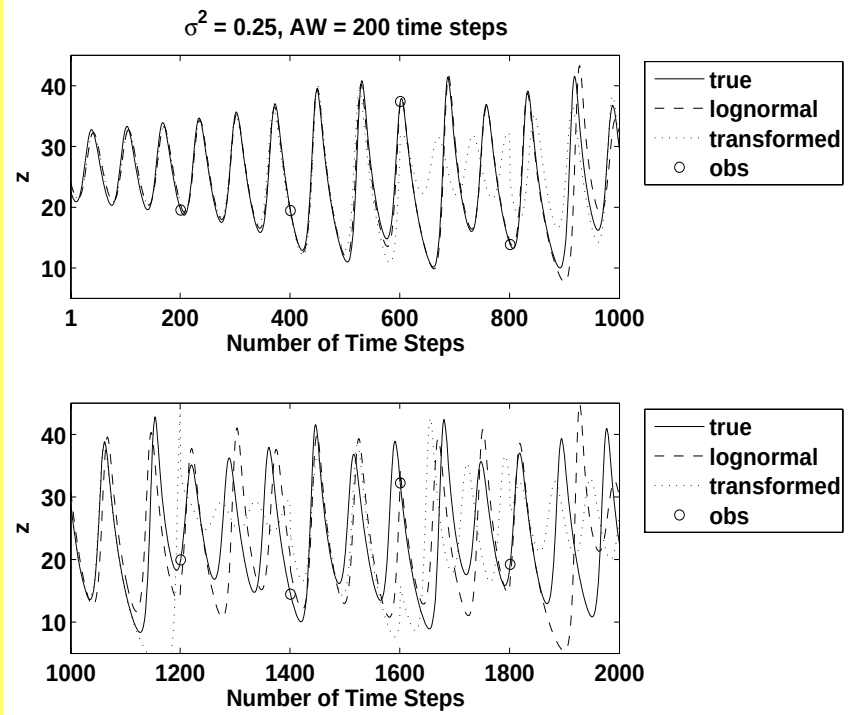




CIRA

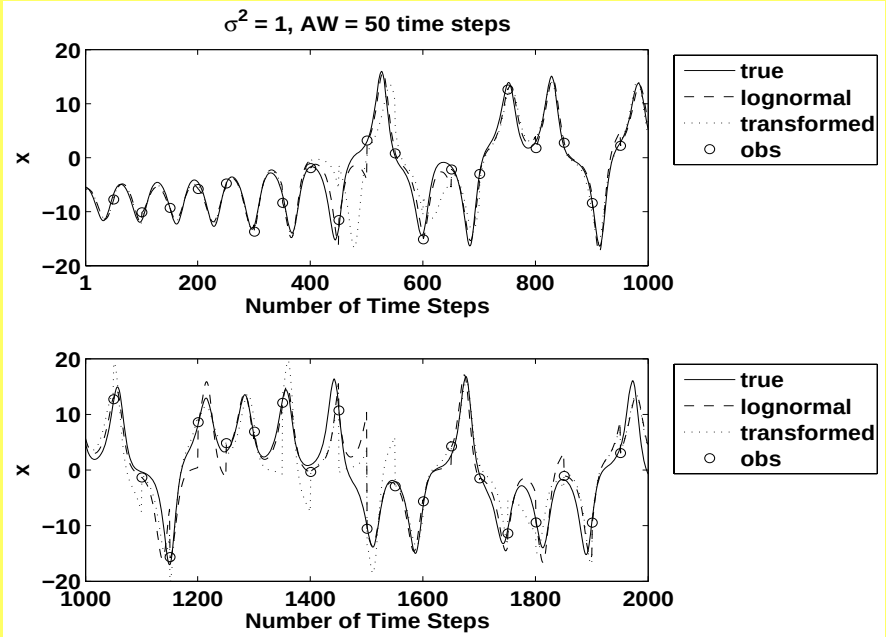
eric Research
tate University

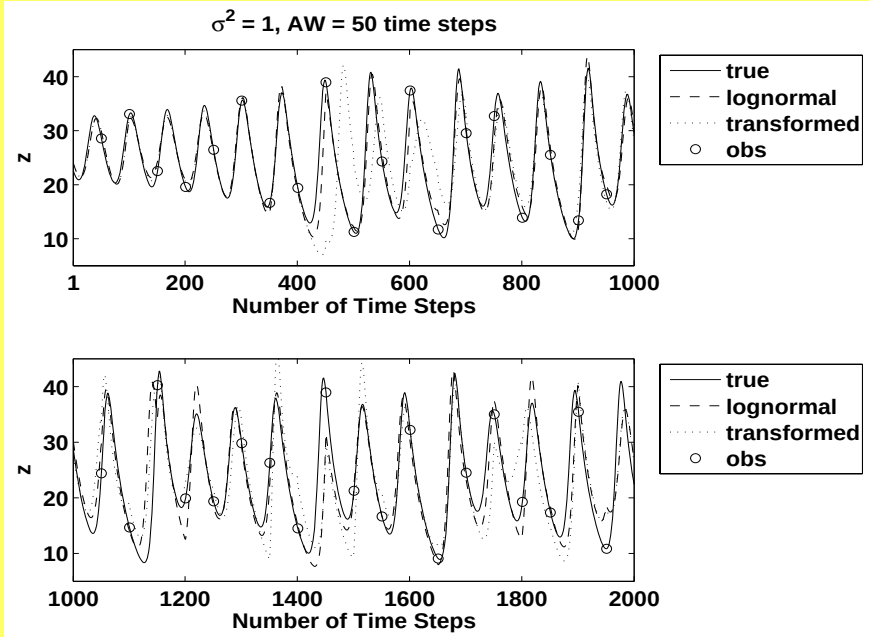


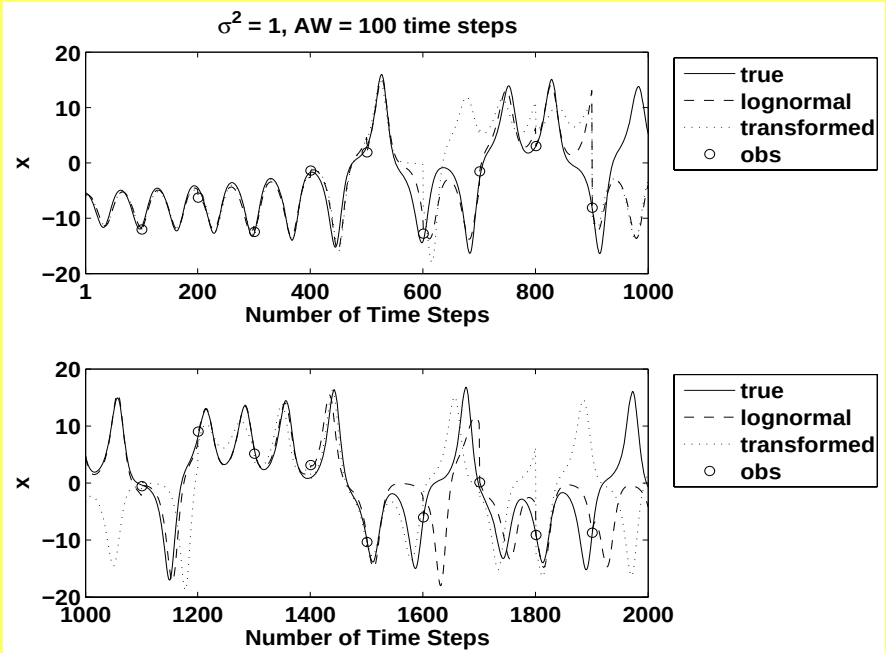


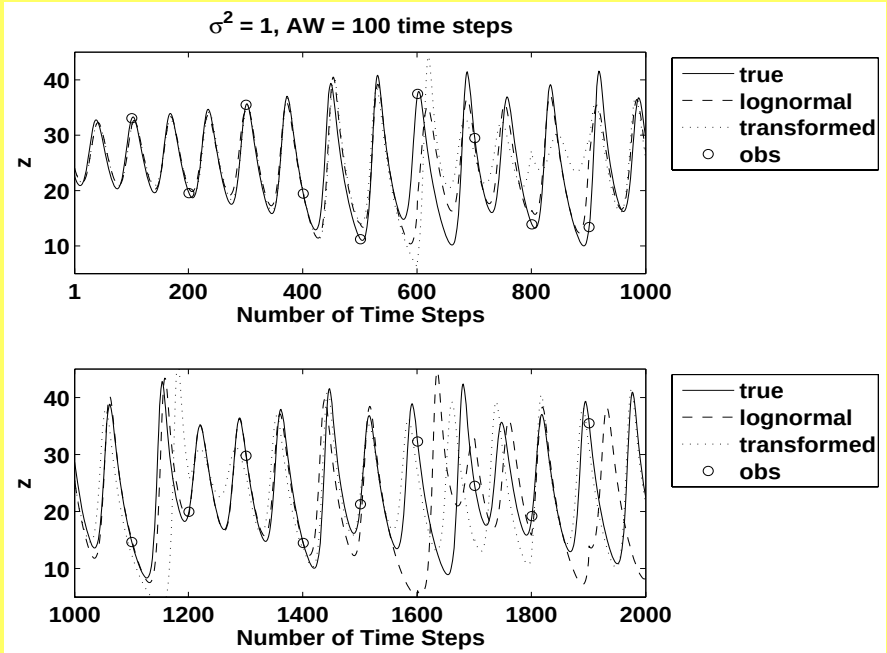
Colorado State University CIRA

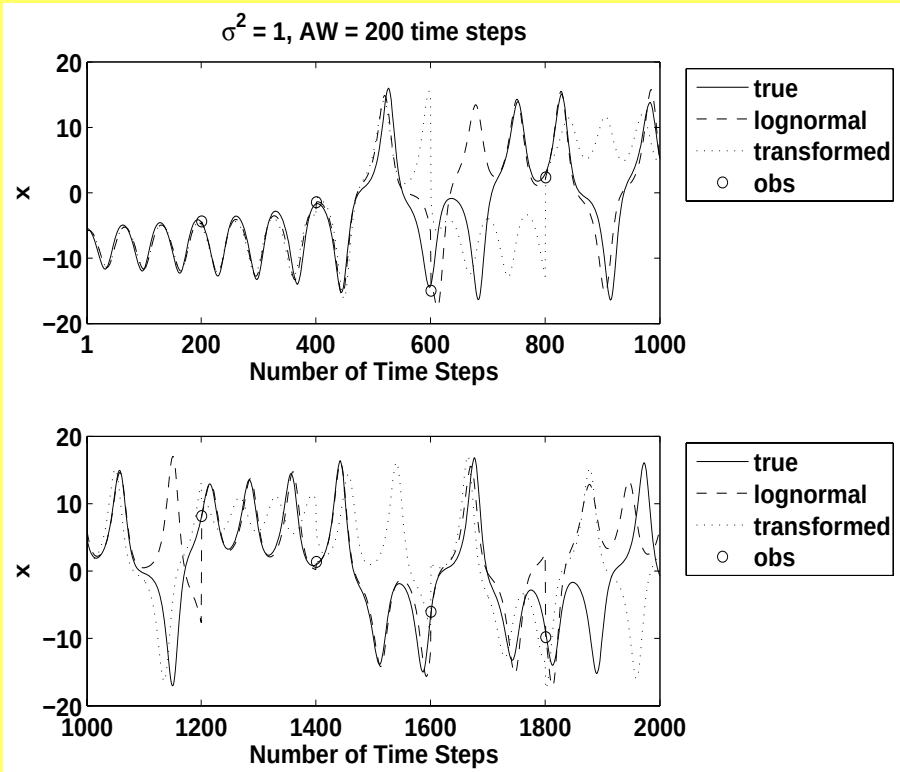
CGIAR DoD Center for Geosciences/Atmospheric Research Colorado State University

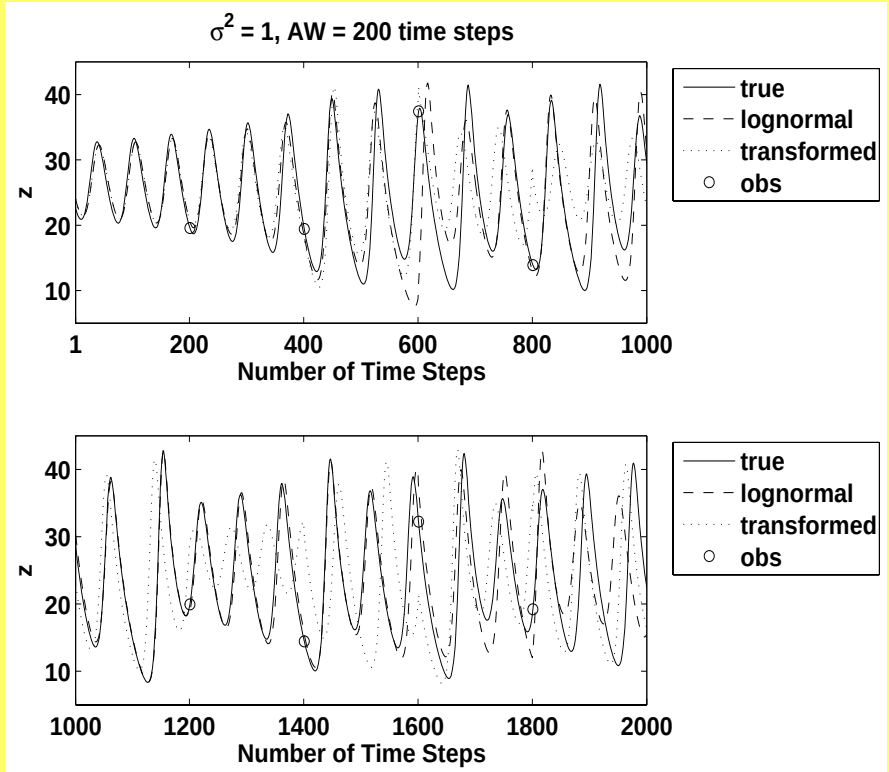

σ² = 1, AW = 50 time steps
x
Number of Time Steps
true
lognormal
transformed
obs


σ² = 1, AW = 50 time steps
z
Number of Time Steps
true
lognormal
transformed
obs


σ² = 1, AW = 100 time steps
x
Number of Time Steps
true
lognormal
transformed
obs


σ² = 1, AW = 100 time steps
z
Number of Time Steps
true
lognormal
transformed
obs





Colorado State University CIRA

CGIAR DoD Center for Geosciences/Atmospheric Research Colorado State University

APPLICATION TO SATELLITE DATA ASSIMILATION

The cost functions which we have presented enable us to avoid the bias correction which is applied in both direct radiance data assimilation and in retrievals.

Given this lognormal framework we can quantify the bias corrections which are currently applied.

TRANSFORMATION (RETRIEVALS)

By transforming to the logarithm of the state variable and the observations which are a mixture then the associated cost function is the standard Gaussian cost function,

$$J(\tilde{x}) = \frac{1}{2}(\tilde{x} - \tilde{x}_b)^T B^{-1}(\tilde{x} - \tilde{x}_b) + \frac{1}{2}(\tilde{y} - \tilde{h}(\tilde{x}))^T R^{-1}(\tilde{y} - \tilde{h}(\tilde{x}))$$

We should have the following cost function

$$J(x) = \frac{1}{2}\hat{\varepsilon}_b^T B^{-1}\hat{\varepsilon}_b + \frac{1}{2}\hat{\varepsilon}_o^T R^{-1}\hat{\varepsilon}_o + \sum_{i=p_1+1}^{N}\hat{\varepsilon}_{bi} + \sum_{j=p_2+1}^{N_o}\hat{\varepsilon}_{oj}$$

SOLUTIONS

The non-linear solutions to the two cost functions on the previous slide are derived in Fletcher and Zupanski (2007) and Lorenc (1986) but are stated here as

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}_b + \boldsymbol{B}\tilde{\boldsymbol{H}}^{\boldsymbol{T}}\tilde{\boldsymbol{W}}_o^T \boldsymbol{R}^{-1}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{h}}(\tilde{\boldsymbol{x}})\right) \tag{6}$$

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_b - \boldsymbol{B}\left[\begin{pmatrix}\boldsymbol{0}_{p_1}\\ \boldsymbol{I}_{q_1}\end{pmatrix} - \boldsymbol{W}_b^{-1}\boldsymbol{H}^T\boldsymbol{W}_o^T\boldsymbol{R}^{-1}\left(\hat{\boldsymbol{\varepsilon}}_o + \boldsymbol{R}\begin{pmatrix}\boldsymbol{0}_{p_2}\\ \boldsymbol{I}_{q_2}\end{pmatrix}\right)\right] \tag{7}$$

$$\boldsymbol{W}_b = \frac{\partial \hat{\boldsymbol{\varepsilon}}_b}{\partial \boldsymbol{x}}, \quad \boldsymbol{W}_o = \frac{\partial \hat{\boldsymbol{\varepsilon}}_o}{\partial \boldsymbol{h}}, \quad \tilde{\boldsymbol{W}}_o = \frac{\partial \tilde{\boldsymbol{\varepsilon}}_o}{\partial \tilde{\boldsymbol{h}}}, \quad \boldsymbol{H} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} \text{ and } \tilde{\boldsymbol{H}} = \frac{\partial \tilde{\boldsymbol{h}}}{\partial \tilde{\boldsymbol{x}}}$$

WRONG DISTRIBUTION

By assuming the wrong distribution we have to see how much of an effect this has. If we assume a Normal background and a set of lognormal observations then if we apply a Taylor series expansion of ε_o then we obtain the following cost function

$$J(x) = \frac{1}{2}(x - x_b)^T B^{-1}(x - x_b)$$

$$+ \frac{1}{2}(y - h(x))^T W_o^T \overline{R}^{-1} W_o (y - h(x)) \quad (8)$$

Where the solution to (8) is given by

$$x = x_b + BH^T W_o^T \overline{R}^{-1} W_o (y - h(x))$$

4D LOGNORMAL DATA ASSIMILATION

Unlike with the three dimensional version of variational data assimilation there is not probability model set out for 4D VAR. Until Fletcher 2007, where we are able to define 4D VAR as a Bayesian Network.

The Gaussian 4D VAR is defined through a calculus of variation problem with initial conditions found through the adjoint. We can do this as well for the lognormal through the inner product

$$g_1(\boldsymbol{x}_0) = \iiint_A \sum_{i=1}^{N_o} \frac{1}{2} \left\langle \ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right), \boldsymbol{R}_i^{-1}\left(\ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right)\right)\right\rangle$$

As with the Gaussian case we know that the first variation of the functional defined on the previous is equivalent to

$$\delta g_1(\boldsymbol{x}_0) = \sum_{i=1}^{N_0} \left\langle \ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right), \boldsymbol{W}_{o,i}\boldsymbol{H}_i \mathbf{M}_i \boldsymbol{R}_i^{-1} \delta \boldsymbol{x}_0 \right\rangle$$

$$= \left\langle \nabla g_1(\boldsymbol{x}_0), \delta \boldsymbol{x}_0 \right\rangle$$

Through using the properties of inner products we get that the gradient is

$$\nabla g_1(\boldsymbol{x}_0) = \sum_{i=1}^{N_o} \left(\boldsymbol{W}_{o,i}\boldsymbol{H}_i \mathbf{M}_i \boldsymbol{R}_i^{-1}\right)^T \left(\ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right)\right)$$

What is wrong with the set up on the previous slide?

The solution is a median and not the mode and hence is independent of the variance.

We need to define the functional as

$$g_2(\boldsymbol{x_0}) =$$

$$\iiint_A \sum_{i=1}^{N_o} \frac{1}{2} \left\langle \ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right) + \boldsymbol{R}^T \boldsymbol{1}, \boldsymbol{R}_i^{-1}\left(\ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right)\right)\right\rangle$$

Which then has a gradient of

$$\nabla g_2(\boldsymbol{x_0}) = \sum_{i=1}^{N_o} \left(\boldsymbol{W}_{o,i} \boldsymbol{H}_i \mathbf{M}_i \boldsymbol{R}_i^{-1}\right)^T \left(\ln \boldsymbol{y}_i - \ln \boldsymbol{h}_i\left(\boldsymbol{M}_i(\boldsymbol{x}_0)\right) + \boldsymbol{R}_i^T \boldsymbol{1}\right)$$

We are only able to define these different functionals because of our previous work with the 3D VAR lognormal data assimilation method and the properties of the three distribution estimators. However, if we wanted to extend this to variables that are not normal or lognormally distributed we need a probability model.

In Fletcher 2007 we present a model through using Bayesian networks theory. The main point about this approach is that it allows us to simplify the expression for Bayes Theorem extended to multiple events which is given by

$$P\left(x_0, x_1, x_2, \ldots, x_{N_o} \middle| y_1, y_2, y_3, \ldots, y_{N_o}\right) =$$
$$P(x_0)P(x_1|x_0)P(y_1|x_1, x_0)P(x_2|y_1, x_1, x_0)$$
$$\ldots P\left(y_{NO} \middle| x_N, y_{N_o-1}, x_{N_o}, \ldots, y_1, x_1, x_0\right)$$

Bayesian networks allow us to remove terms that are not conditioned on other random variables. From the diagram on the board for the perfect model we have that all the model states are only dependent on the initial conditions. We also have that the observations are only dependent on the model state at the time which allows further simplification to give us

$$P\left(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x}_{N_o} \middle| \boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \ldots, \boldsymbol{y}_{N_o}\right) =$$

$$P\left(\boldsymbol{x_0}\right) \prod_{i=1}^{N_o} P\left(\boldsymbol{y}_i \middle| \boldsymbol{x_0}\right)$$

For the multivariate Gaussian case we have

$$P\left(\boldsymbol{x_0}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{x_0} - \boldsymbol{x}_{b,0}\right)^T \boldsymbol{B}_0^{-1}\left(\boldsymbol{x_0} - \boldsymbol{x}_{b,0}\right)\right\}$$

$$P\left(\boldsymbol{y}_i \middle| \boldsymbol{x_0}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{y_i} - \boldsymbol{h}_i\left(\boldsymbol{M}_i\left(\boldsymbol{x_0}\right)\right)\right)^T \boldsymbol{R}_i^{-1}\left(\boldsymbol{y_i} - \boldsymbol{h}_i\left(\boldsymbol{M}_i\left(\boldsymbol{x_0}\right)\right)\right)\right\}$$

For the multivariate lognormal case we have

$$P(\boldsymbol{x_0}) \propto \left(\prod_{j=1}^{N} \frac{\boldsymbol{x_{0,i}}}{\boldsymbol{x_{b,0,j}}} \right) \times$$

$$\exp\left\{ -\frac{1}{2} \left(\ln \boldsymbol{x_0} - \ln \boldsymbol{x_{b,0}} \right)^T \boldsymbol{B}_0^{-1} \left(\ln \boldsymbol{x_0} - \ln \boldsymbol{x_{b,0}} \right) \right\}$$

$$P(\boldsymbol{y}_i | \boldsymbol{x_0}) \propto \left(\prod_{k=1}^{N_{o,i}} \frac{h_{i,k}(\boldsymbol{M}(\boldsymbol{x}_0))}{y_{i,k}} \right) \times$$

$$\exp\left\{ -\frac{1}{2} \left(\boldsymbol{y_i} - \boldsymbol{h}_i(\boldsymbol{M}_i(\boldsymbol{x_0})) \right)^T \boldsymbol{R}_i^{-1} \left(\boldsymbol{y_i} - \boldsymbol{h}_i(\boldsymbol{M}_i(\boldsymbol{x_0})) \right) \right\}$$

REFERENCES

COHN, S.E., 1997: AN INTRODUCTION TO ESTIMATOR ERROR THEORY. *J. Met. Soc. Japan,* **75**, 420-436.

DERBER, J.C. AND WU, W.-S., 1998: THE USE OF TOVS CLOUD-CLEARED RADIANCES IN THE NCEP SSI ANALYSIS SYSTEM. *Mon. Wea. Rev.,* **126,** 2287-2299.

EVENSON, G., 1994: SEQUENTIAL DATA ASSIMILATION WITH A NONLINEAR QUASI-GEOSTROPHIC MODEL USING MONTE-CARLO METHODS TO FORECAST ERROR STATISTICS. *J. Geophys. Res.* **99** (C5), 10,143-10,162.

FLETCHER, S.J. AND ZUPANSKI, M. 2006a: A DATA ASSIMILATION METHOD FOR LOGNORMALLY DISTRIBUTED OBSERVATIONAL ERRORS. In Print: *Q. J. R. Meteor. Soc*.

FLETCHER, S.J. AND ZUPANSKI, M. 2006b: A HYBRID MULTIVARIATE NORMAL AND LOGNORMAL `DISTRIBUTION FOR DATA ASSIMILATION. *Atmos. Sci. Letters.* **7**, 43-46.

FLETCHER, S.J. AND ZUPANSKI, M. 2007: A CAVEAT CONCERNING THE USE OF NORMAL FRAMEWORKS FOR LOGNORMAL VARIABLES AND OBSERVATIONS IN VARIATIONAL DATA ASSIMILATION. Submitted *Mon. Wea. Rev.*

HARRIS, B.A. AND KELLY, G. 2001: A SATELLITE RADIANCE-BIAS CORRECTION SCHEME FOR DATA ASSIMILATION. *Q. J. Roy. Meteor. Soc.* **127,** 1453-1486.

HEYDE, C.C., 1963: ON A PROPERTY OF THE LOGNORMAL DISTRIBUTION. *J. Roy. Stats. Soc. Ser. B.,* **25**, 392-393.

KALMAN, R.E., 1960: A NEW APPROACH TO LINEAR FILTERING AND REDICTION PROBLEMS. *Trans. ASME. J. Basic. Eng.,* **82,** 35-45.

LORENC, A.C., 1986: ANALYSIS METHODS FOR NUMERICAL WEATHER PREDICTION. *Q. J. Roy. Meteor. Soc.* **112**, 1177-1194.

REFERENCES

MIELKE Jr., P.W., WILLIAMS, J.S. AND WU, S.-U., 1997: COVARIANCE ANALYSIS TECHNIQUES BASED UPON BIVARAITE LOG-NORMAL DISTRIBUTION WITH WEATHER MODIFIATION APPLICATION. *J. Appl. Meteorol.,* **16,** 183-187.

MILES N.L., VERLINDE, J. AND CLOTHIAUX, E.E., 2000: CLOUD DROPLET SIZE DISTRIBUTION IN LOW-LEVEL STRATIFORM CLOUDS.*J. Atmos. Sci,* **57,** 295-311.

POLAVARAPU, S., REN, S., ROCHON, Y., SANKEY, D., EK, N., KOSHYK, J. AND TARASICK, D., 2005: DATA ASSIMILATION WITH THE CANDIAN MIDDLE ATMOSPHERE MODEL. *Atmosphere-Ocean* **43(1),** 77-100.

RAYMOND, W.H., 1997: A THEORETICAL EVALUATION OF THE REVELENCE OF LOGNORMAL DISTRIBUTIONS FOR THE MOISTURE FLUX AND WIND COMPONENTS. *Mon. Wea. Rev.,* **125,** 3018-3023

STEPHENS, G.L. AND COAUTHORS, 2002: THE CLOUDSAT MISSION AND THE A-TRAIN. *Bull. Amer. Meteor. Soc.* **83,** 1771-1190.

MINIMISATION AND PRECONDITIONING

Outline of lecture

1) One dimension Newton-Rhapson

2) Unconstrained optimisation

3) Newton's method for non-linear minimisation

4) Wolfe Condition

5) Quasi-Newton methods

6) Conjugate Gradients

7) Preconditioning

8) Control Variable Transforms

Newton's method for one-dimensional problems

We wish to solve the problem f(x)=0 but we can not analytically find the solutions so we approximate the gradient with a tangent line to the function given a good guess to the true solution. We obtain the iterative formula from a Taylor series expansion of the function in the vicinity of the true root.

$$f(x) = f(x_n) + (x - x_n) f'(x_n) + \frac{(x - x_n)^2}{2} f''(\xi)$$

$$x \le \xi \le x_n. \text{ Let } x = \alpha \Rightarrow f(\alpha) = 0$$

$$\therefore \alpha = x_n - \frac{f(x_n)}{f'(x_n)}$$

Newton's method in multiple dimensions

Like with one dimension case the multi-dimensional version of Newton's method come from a Taylor series which then gives

$$x_{n+1} = x_n - F^{-1}(x_n) f(x_n)$$

$$F^{-1}(x_n) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1} & \frac{\partial f_N}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_N} \end{pmatrix}$$

UNCONSTRAINED OPTIMISATION

In this class of problems we are seeking the minimum/maximum of a continuous function of several variables. To constrain the solution we have that a point α is called a STRICT LOCAL MINIMUM of the function *f* if

$$f(\boldsymbol{x}) > f(\boldsymbol{\alpha}) \quad \forall \boldsymbol{x} \text{ close to } \boldsymbol{\alpha} \text{ and } \boldsymbol{x} \neq \boldsymbol{\alpha}$$

Generally an initial guess of α will be well known and we also assume that the function is twice differentiable with respect to the variables.

We now reformulate Newton's method for the function to use the famous calculus result that at the minima then the gradient of the function is zero, i.e.

$$\frac{\partial f(\boldsymbol{\alpha})}{\partial x_i} = 0, \quad i = 1,2,\ldots,N$$

Thus we are having to solve the non-linear system of equations given by

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} = 0, \quad i = 1,2,\ldots,N$$

Which in vector notation this problem looks like

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_N} \right)^T = \boldsymbol{0}$$

To formulate the minimisation scheme we consider a Taylor series expansion of the gradient of the function.

$$0 = \nabla f(\alpha) \approx \nabla f(x_n) + (x_{n+1} - x_n) H(x_n)$$

$$-\nabla f(x_n) = (x_{n+1} - x_n) H(x_n)$$

$$(x_{n+1} - x_n) = -H^{-1}(x_n) \nabla f(x_n)$$

$$x_{n+1} = x_n - H^{-1}(x_n) \nabla f(x_n)$$

Where

$$H(x)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad 1 \le i, j \le N$$

Is the Hessian matrix. The problem for the weather prediction community is that the size of this matrix is 10^7 by 10^7 and therefore we can not analytically find G. Therefore we need other methods. These are called the descent methods.

At the point $\boldsymbol{x}_n$ pick a direction $\boldsymbol{d}_n$ such that $f(\boldsymbol{x}_n)$ will decrease as $\boldsymbol{x}$ moves away from $\boldsymbol{x}_{\mathrm{n}}$ in the direction, $\boldsymbol{d}_n$.

We start by letting

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + s\boldsymbol{d}_n$$

Usually s is choosen as the smallest positive relative minimum of $\phi(\mathrm{s})$ therefore with each iteration

$$f(\boldsymbol{x}_{n+1}) < f(\boldsymbol{x}_n)$$

How do we choose $\boldsymbol{d_n}$?

1) The method of steepest descent uses

$$\boldsymbol{d_n = -\nabla f(x_n)}$$

However this is not known for fast convergence.

2) Quasi - Newton methods are such that approximations are made to the Hessian matrix.

However we need a condition of the term s. These are the Wolfe conditions

Wolfe Conditions

Let $f : \Re^N \to \Re$ be a smooth objective function, and let $\boldsymbol{d}_n$ be a given search direction.

A step length s_n is said to satisfy the Wolfe conditions if

$i)\ f(\boldsymbol{x}_n + s_n \boldsymbol{d}_n) \le f(\boldsymbol{x}_n) + c_1 s_n \boldsymbol{d}_n^T \nabla f(\boldsymbol{x}_n)$ (Armijo Condition)

$ii)\ \boldsymbol{d}_n^T \nabla f(\boldsymbol{x}_n + s_n \boldsymbol{d}_n) \ge c_2 \boldsymbol{d}_n^T \nabla f(\boldsymbol{x}_n)$ (Curvature Condition)

There are problems with the second condition and so is made stronger by the following adjustment

$$\left| \boldsymbol{d}_n^T \nabla f(\boldsymbol{x}_n + s_n \boldsymbol{d}_n) \right| \le c_2 \left| \boldsymbol{d}_n^T \nabla f(\boldsymbol{x}_n) \right|$$

$i)$ combined with the new condition above is refered to as the strong Wolfe conditions .

Quasi-Newton Methods

step i) : $\Delta \boldsymbol{x}_{\boldsymbol{n}} = -s_n H_n^{-1} \nabla f\left(\boldsymbol{x}_{\boldsymbol{n}}\right)$ where s_n satisfies the Strong Wolfe Conditions

step ii) : $\boldsymbol{x}_{\boldsymbol{n+1}} = \boldsymbol{x}_{\boldsymbol{n}} + \Delta \boldsymbol{x}_{\boldsymbol{n}}$

step iii) : Calculate the new gradient $\nabla f\left(\boldsymbol{x}_{\boldsymbol{n+1}}\right)$ at the new point.

This then enables us to update the Hessian by the approximation

$$\boldsymbol{y}_n = \nabla f\left(\boldsymbol{x}_{\boldsymbol{n+1}}\right) - \nabla f\left(\boldsymbol{x}_{\boldsymbol{n}}\right)$$

However, this is a very basic approximation. Therefore this has been much research into different approximations to the Hessian matrix. One very well used approach is the Broyden - Fletcher - Goldfarb - Shanno (BFGS) method.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) Method

The limited memory version of this algorithm is used in many data assimilation methods. This approach is used in NCEP's Meso-scale 4D VAR system, (ETA VAR), it is also used in the Cooperative Institute for Research in the Atmosphere (CIRA)/Colorado State University's 4D VAR system, RAMDAS – Regional Atmospheric Modeling Data Assimilation System, Zupanski *et al.* (2005). This is also used in the United Kingdom's Met Office 4D VAR operational weather prediction systems from synoptic to cloud resolving data assimilation systems.

The Hessian matrix is approximated by the addition of two more matrices. The algorithm starts from an initial guess for the true state $\boldsymbol{x}_0$ and the Hessian matrix $\boldsymbol{B}_0$. Then the algorithm is given by the following 4 steps.

step 1) Obtain $\boldsymbol{s}_n$ by solving $\boldsymbol{B}_n \boldsymbol{s}_n = -\nabla \boldsymbol{f}(\boldsymbol{x}_n)$.

step 2) Perform a line search to find the optimal α_{n} in the direction found in the first step then update the state

$$\boldsymbol{x}_{\mathrm{n+1}} = \boldsymbol{x}_{\mathrm{n}} + \alpha_{\mathrm{n}} \boldsymbol{s}_{\mathrm{n}}$$

Step 3) $\boldsymbol{y}_{\mathrm{n}} = \nabla \boldsymbol{f}(\boldsymbol{x}_{n+1}) - \nabla \boldsymbol{f}(\boldsymbol{x}_n)$

Step 4) $$\boldsymbol{B}_{n+1} = \boldsymbol{B}_n + \frac{\left(\boldsymbol{y}_n \boldsymbol{y}_n^T\right)}{\left(\boldsymbol{y}_n^T \boldsymbol{s}_n\right)} - \frac{\left(\boldsymbol{B}_n \boldsymbol{s}_n \boldsymbol{s}_n^T \boldsymbol{B}_n^T\right)}{\left(\boldsymbol{s}_n^T \boldsymbol{B}_n \boldsymbol{s}_n\right)}$$

Non-linear Conjugate Gradients methods

These approaches are used instead of Quasi-Newton methods as they avoid the Hessian approximations but are not as fast to converge as the Newton methods. These approaches are seeking solutions to the problem

$$f(\boldsymbol{x}) = |\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}|^2$$

Where the minimum occurs when the gradient is zero i.e.

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = 0$$

The non-linear CG methods seek a solution to the gradient problem same as the Newton methods.

So given a cts function to minimise the gradient indicates the direction of a maximum increment. Therefore we start in the opposite (steepest descent) direction.

$$1)\Delta \boldsymbol{x}_n = -\nabla f\left(\boldsymbol{x}_n\right)$$

2) compute the scalar β_{n} from the following two choices

$$\beta_{\mathrm{n}}^{\mathrm{FR}} = \frac{\Delta \boldsymbol{x}_n^T \Delta \boldsymbol{x}_n}{\Delta \boldsymbol{x}_{n-1}^T \Delta \boldsymbol{x}_{n-1}}$$

$$\beta_{\mathrm{n}}^{\mathrm{PR}} = \frac{\Delta \boldsymbol{x}_n^T \left(\Delta \boldsymbol{x}_n - \Delta \boldsymbol{x}_{n-1}\right)}{\Delta \boldsymbol{x}_{n-1}^T \Delta \boldsymbol{x}_{n-1}}$$

$$3)\Delta \boldsymbol{x}_n = \Delta \boldsymbol{x}_n + \beta_{\mathrm{n}} \Delta \boldsymbol{x}_{n-1}$$

4) Optimise α_n min $f\left(\boldsymbol{x}_n + \alpha_n \Delta \boldsymbol{x}_n\right)$

$$5)\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \alpha_n \Delta \boldsymbol{x}_n$$

Minimization in ensemble DA

Direct solution of Extended Kalman Filter:

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{P}_f \boldsymbol{K}^T \left(\boldsymbol{K} \boldsymbol{P}_f \boldsymbol{K}^T + \boldsymbol{R} \right)^{-1} [\boldsymbol{y} - K(\boldsymbol{x}_b)]$$

$$\boldsymbol{P}_a = \boldsymbol{P}_f - \boldsymbol{P}_f \boldsymbol{K}^T (\boldsymbol{K} \boldsymbol{P}_f \boldsymbol{K}^T + \boldsymbol{R})^{-1} \boldsymbol{K} \boldsymbol{P}_f$$

Equivalent to solving a *quadratic* minimization problem (Lorenc 1986):

$$J = \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_b)^T \boldsymbol{P}_f^{-1} (\boldsymbol{x} - \boldsymbol{x}_b) + \frac{1}{2} [\boldsymbol{y} - K(\boldsymbol{x})]^T \boldsymbol{R}^{-1} [\boldsymbol{y} - K(\boldsymbol{x})]$$

Subject to $K(\boldsymbol{x}) = K(\boldsymbol{x}_b) + \boldsymbol{K}(\boldsymbol{x} - \boldsymbol{x}_b)$

$$\boldsymbol{K} = \left(\frac{\partial K}{\partial \boldsymbol{x}} \right)$$

Ensemble DA is a minimization process. Due to perfect Hessian preconditioning, the minimum solution is obtained in a single minimization iteration (for near-quadratic problem).

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

Minimization in variational DA

Variational DA is a minimization algorithm

$$J = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_b)^T \boldsymbol{P}_f^{-1}(\boldsymbol{x} - \boldsymbol{x}_b) + \frac{1}{2}[\boldsymbol{y} - K(\boldsymbol{x})]^T \boldsymbol{R}^{-1}[\boldsymbol{y} - K(\boldsymbol{x})]$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{P}^{-1}\boldsymbol{g}_k$$

$\boldsymbol{P}$ - preconditioning matrix

- Gradient based minimization (CG, quasi-Newton, truncated Newton)
  - ❑ Adjoint
  - ❑ Multiple tangent linear models
  - ❑ Finite differences

- Difficult preconditioning due to large dimensions of the problem

- Preconditioning impacts not only the minimization convergence (cost), but also the accuracy of the solution (quality)

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

Hessian preconditioning

Hessian and inverse Hessian (linear contribution)

$$\boldsymbol{H} = \frac{\partial^2 J}{\partial \boldsymbol{x}^2} = \boldsymbol{P}_f^{-1} + \boldsymbol{K}^T \boldsymbol{R}^{-1} \boldsymbol{K} = \boldsymbol{P}_f^{-\frac{T}{2}} (\boldsymbol{I} + \boldsymbol{A}) \boldsymbol{P}_f^{-\frac{1}{2}} \qquad \boldsymbol{H} = \boldsymbol{E}\boldsymbol{E}^T$$

$$\boldsymbol{H}^{-1} = \left(\frac{\partial^2 J}{\partial \boldsymbol{x}^2} \right)^{-1} = \boldsymbol{P}_f^{1/2} (\boldsymbol{I} + \boldsymbol{A})^{-1} \boldsymbol{P}_f^{T/2}$$

$$\boldsymbol{A} = \boldsymbol{P}_f^{T/2} \boldsymbol{K}^T \boldsymbol{R}^{-1} \boldsymbol{K} \boldsymbol{P}_f^{1/2}$$

Change of variable (preconditioning):

$$\boldsymbol{x} - \boldsymbol{x}_B = \boldsymbol{E}^{-T} \boldsymbol{\zeta} = \boldsymbol{P}_f^{1/2} (\boldsymbol{I} + \boldsymbol{A})^{-T/2} \boldsymbol{\zeta}$$

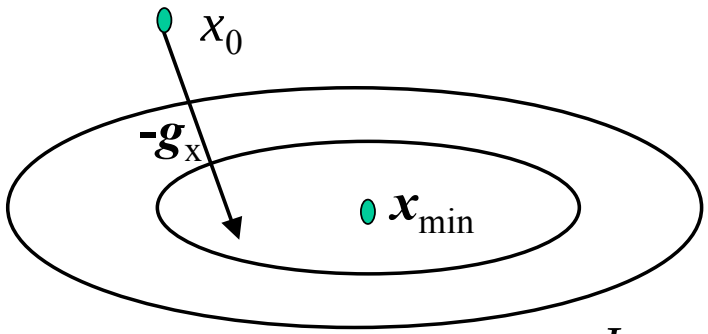
$$\boldsymbol{H}_\zeta = \boldsymbol{E}^{-1} \boldsymbol{H} \boldsymbol{E}^{-T} = \boldsymbol{E}^{-1} \boldsymbol{E} \boldsymbol{E}^T \boldsymbol{E}^{-T} = \boldsymbol{I}$$

Ideal Hessian preconditioning

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

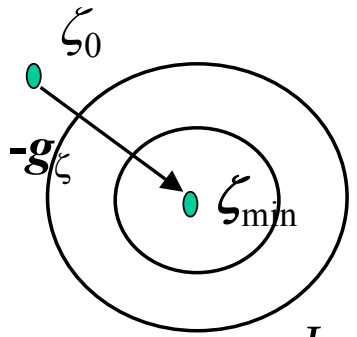
Ideal Hessian Preconditioning

***Physical space* (*x*)**



$J=const.$

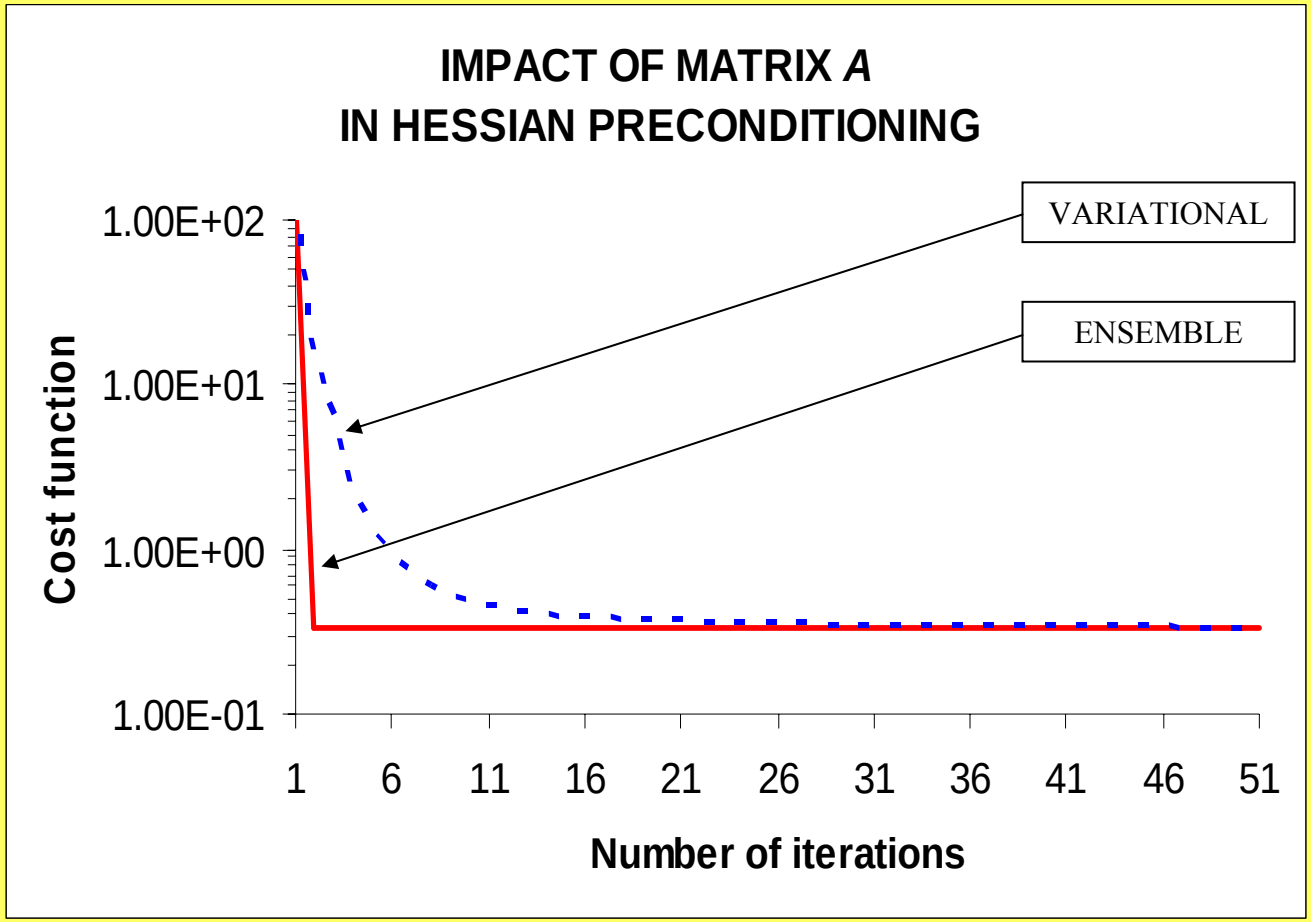
***Preconditioning space* (ζ)**



$J=const.$

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

Hessian Preconditioning



$$\boldsymbol{P}_{VAR}^{-1} = \boldsymbol{P}_f$$

$$\boldsymbol{P}_{ENS}^{-1} = \boldsymbol{P}_f^{1/2}(\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{P}_f^{T/2}$$

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

Impact of preconditioning in NCEP's Eta 4DVAR system

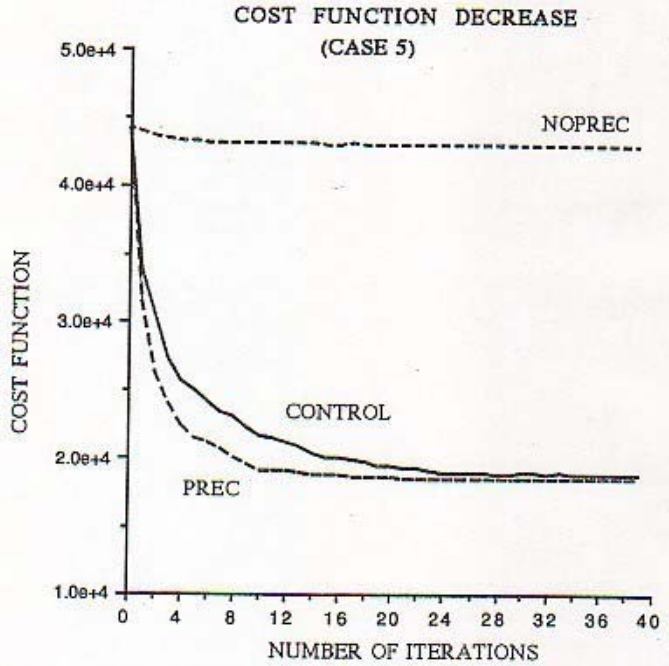


FIG. 2. The cost function decrease during the first 40 iterations of minimization algorithm, for the assimilation period 0000–1200 UTC 24 November 1994 (case 5). The three experiments shown include CONTROL, with background variance used as a preconditioner (full line); NOPREC, with no preconditioning used (short-dashed line); and PREC, with the proposed preconditioning method used (long-dashed line). (Zupanski 1996, Mon. Wea. Rev.)

Incorrect solution without preconditioning !

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

PRECONDITIONING IN 4DVAR (NCEP ETA, CSU RAMS)

- Controls the magnitude of control variable adjustment

$$d = P^{-1} f(g)$$

Synoptic situation dependent (D=diagonal empirical matrix)

$$\frac{\partial^2 J}{\partial x^2} = B^{1/2}(I + D)^{-1} B^{T/2}$$

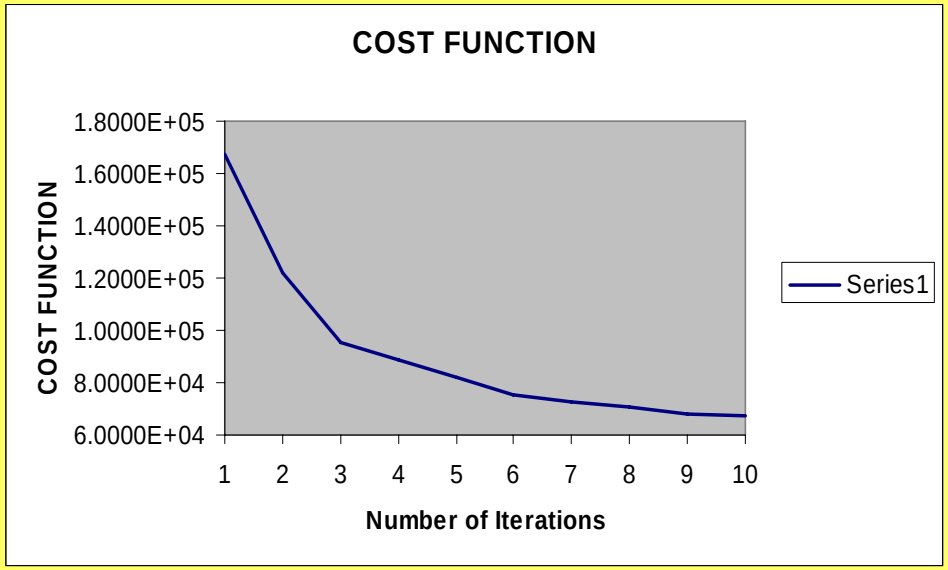
4DVAR RESULTS WITH *RAMS* MODEL

- **6-HOUR ASSIMILATION PERIOD**: 03/08/2002 12 Z – 18 Z
- **WRF observations** at 15Z and 18Z
- **Control Variables**: *INITIAL CONDITIONS + MODEL ERROR*
- *- perturbation Exner function*
- *- potential temperature*
- *- horizontal winds*
- *- total mixing ratio*
- **4DVAR analysis**: ***END*** of the assimilation interval
- **CSU RAMS** non-hydrostatic NWP model, Level 2 Microphysics
- **Adjoint** with explicit microphysics
- 15 km horizontal resolution, 31 vertical level
- 120 X 80 grid points (1800 km X 1200 km)
- **Calculations** performed on Linux PC Cluster (8 CPU)
- **10 MINIMIZATION ITERATIONS**
- **ONE** minimization iteration cost ~ 10 RAMS model integrations (6-h forecast)

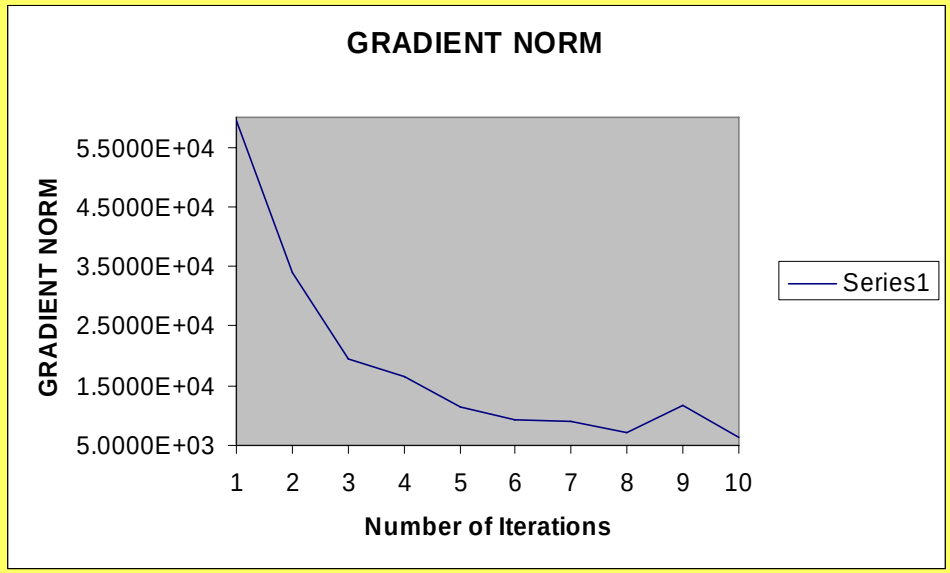
Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

MINIMIZATION ITERATIONS



Smooth Convergence



Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

COST FUNCTION

Cost Function

$$J(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_B)^T \boldsymbol{P}_f^{-1}(\boldsymbol{x}-\boldsymbol{x}_B) + \frac{1}{2}\sum_{n=1}^{N}[K(M(\boldsymbol{x})) - \boldsymbol{y}]_n^T \boldsymbol{R}_n^{-1}[K(M(\boldsymbol{x}) - \boldsymbol{y}]_n$$

$\boldsymbol{x}$ – ***control variable*** ~ O(10^8)

n – time index (n=1, . . ., N)
K – (non-linear) observation operator
M – (non-linear) NWP model
$\boldsymbol{y}$ – observation vector
$\boldsymbol{R}$ – observation error covariance
$\boldsymbol{x_B}$ – first-guess (background) model state
$\boldsymbol{P}_f$ – forecast error covariance

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

HESSIAN

Hessian matrix (linear contribution)

$$\boldsymbol{H} = \frac{\partial^2 J}{\partial \boldsymbol{x}^2} = \boldsymbol{P}_f^{-1} + \sum_{n=1}^{N} \boldsymbol{M}^T \boldsymbol{K}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{K}_n \boldsymbol{M} = \boldsymbol{P}_f^{-\frac{T}{2}} (\boldsymbol{I} + \boldsymbol{A}) \boldsymbol{P}_f^{-\frac{1}{2}}$$

$$\boldsymbol{A} = \sum_{n=1}^{N} \boldsymbol{P}_f^{T/2} \boldsymbol{M}^T \boldsymbol{K}_n^T \boldsymbol{R}_n^{-1} \boldsymbol{K}_n \boldsymbol{M} \boldsymbol{P}_f^{1/2}$$

$$\boldsymbol{H} = \boldsymbol{E}\boldsymbol{E}^T$$

Change of variable (preconditioning):

$$\boldsymbol{x} - \boldsymbol{x}_B = \boldsymbol{E}^{-T} \boldsymbol{\zeta} = \boldsymbol{P}_f^{1/2} (\boldsymbol{I} + \boldsymbol{A})^{-T/2} \boldsymbol{\zeta}$$

$$\boldsymbol{H}_\zeta = \boldsymbol{E}^{-1} \boldsymbol{H} \boldsymbol{E}^{-T} = \boldsymbol{E}^{-1} \boldsymbol{E}\boldsymbol{E}^T \boldsymbol{E}^{-T} = \boldsymbol{I}$$

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

PRECONDITIONING *ISSUES* IN APPLICATIONS TO REALISTIC ATMOSPHERIC AND OCEANIC DATA ASSIMILATION

(1) Large dimensions of the Hessian

(2) Unavailable matrix representation of the Hessian

(3) Inversion of 10^8 X 10^8 matrix not computationally feasible

(4) Computationally expensive calculation of the cost-function and gradient

(5) Preconditioning is necessary: Hessian Condition Number ~ $O(10^{15})$

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

PRECONDITIONING *SOLUTIONS* IN APPLICATIONS TO REALISTIC ATMOSPHERIC AND OCEANIC DATA ASSIMILATION

(1) (Square-root) forecast error covariance matrix: $E^{-T} \cong P_f^{1/2}$

- assumption: $A = I$

Hessian Condition Number ~ O(10²)- O(10³)

(2) Empirical preconditioning: $E^{-T} \cong P_f^{1/2}(I + D)^{-\frac{T}{2}}$

- assumption: $A = D$ *D – empirical diagonal matrix*

Hessian Condition Number ~ O(10¹)

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

EMPIRICAL PRECONDITIONING
(Zupanski - Tellus 1993, MWR 1996)

Iterative minimization: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \;\; ; \;\; \boldsymbol{H}\boldsymbol{d}_k = -\boldsymbol{g}_k$

USE TAYLOR EXPANSION OF THE COST FUNCTION:

$$J_{k+1} = J_k + \alpha \boldsymbol{g}^T \boldsymbol{d} + \frac{1}{2}\alpha \boldsymbol{d}^T \boldsymbol{H}\boldsymbol{d}$$

$$\Delta J = J_k - J_{k+1} = \alpha(1 - \frac{\alpha}{2})(-\boldsymbol{g}^T \boldsymbol{d})$$

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

EMPIRICAL PRECONDITIONING – cont.1

NOTE: *Two (or more) components of the cost-function and the inner product*

$$J = J^B + J^{obs} \Rightarrow \Delta J = \Delta J^B + \Delta J^{obs}$$

$$(-\boldsymbol{g}^T \boldsymbol{d}) = (-\boldsymbol{g}^T \boldsymbol{d})^B + (-\boldsymbol{g}^T \boldsymbol{d})^{obs}$$

$$\boldsymbol{H} = \frac{\partial^2 J^B}{\partial \boldsymbol{x}^2} + \frac{\partial^2 J^{obs}}{\partial \boldsymbol{x}^2} = \boldsymbol{P}_f^{-1} + \boldsymbol{P}_f^{-\frac{T}{2}} \boldsymbol{A} \boldsymbol{P}_f^{-\frac{1}{2}}$$

REMARKS:

- *Taylor expansion of the cost function can be applied to the total cost function, or to its components*
- *In this application, the background has well-defined covariance (P_f)*
- *Apply preconditioning approximation only to the observational component of the cost function*

$$\boldsymbol{H}^{obs} = \frac{\partial^2 J^{obs}}{\partial \boldsymbol{x}^2} = \boldsymbol{P}_f^{-\frac{T}{2}} \boldsymbol{A} \boldsymbol{P}_f^{-\frac{1}{2}}$$

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

EMPIRICAL PRECONDITIONING – cont.2

Assumption:

MOST IMPORTANT VARIABILITY OF THE HESSIAN MATRIX IS

(1) In vertical coordinate direction

(2) For different physical variables

$$\Delta J_L = \alpha(1-\frac{\alpha}{2})(-\boldsymbol{g}^T \boldsymbol{d})_L$$

INTRODUCE UNKNOWN DIAGONAL MATRIX $\boldsymbol{D}$:

(1) Assume the Hessian preconditioning in the form

$$(\boldsymbol{E}\boldsymbol{E}^T)^{-1} = \boldsymbol{P}_f^{1/2}\boldsymbol{D}^{-1}\boldsymbol{P}_f^{T/2}$$

(2) Matrix $\boldsymbol{D}$ elements change ONLY in vertical, and for physical variables

(3) NOTE: Only observational component of the cost function considered, the prior has known Hessian

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu

EMPIRICAL PRECONDITIONING – cont.3

Descent direction (unknown D):

$$\boldsymbol{d} = -\boldsymbol{P_f}^{1/2}\boldsymbol{D}^{-1}\boldsymbol{P_f}^{T/2}\boldsymbol{g}$$

Gradient norm:

$$(-\boldsymbol{g}^T\boldsymbol{d})_L = (\boldsymbol{g}^T\boldsymbol{P}_f^{\frac{1}{2}}\boldsymbol{D}^{-1}\boldsymbol{P}_f^{\frac{T}{2}}\boldsymbol{g})_L \cong D_L^{-1}(\boldsymbol{g}^T\boldsymbol{P}_f\boldsymbol{g})_L$$

Cost function:

$$\Delta J_L = \beta\ J_L \qquad \beta < 1$$

Assumed decrease of the cost function (most often $\beta \sim 0.5$)

In common applications, P_f defined over horizontal grid only

$$(\boldsymbol{g}^T\boldsymbol{P}_f^{\frac{1}{2}}\boldsymbol{D}^{-1}\boldsymbol{P}_f^{\frac{T}{2}}\boldsymbol{g})_L = \boldsymbol{g}_L{}^T(\boldsymbol{P}_f^{\frac{1}{2}})_L\boldsymbol{D}_L^{-1}(\boldsymbol{P}_f^{\frac{T}{2}})_L\boldsymbol{g}_L = D_L^{-1}(\boldsymbol{g}^T\boldsymbol{P}_f\boldsymbol{g})_L$$

Colorado State University CIRA

Milija Zupanski, CIRA/CSU
ZupanskiM@CIRA.colostate.edu