

A hybrid multivariate Normal and lognormal distribution for data assimilation

Steven J. Fletcher* and Milija Zupanski

Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, USA

*Correspondence to:

Steven J. Fletcher, Cooperative
Institute for Research in the
Atmosphere, Colorado State
University, 1375 Campus
Delivery, Fort Collins, CO
80523-1375, USA.

E-mail:

fletcher@cira.colostate.edu

Abstract

In this article, we define and prove a distribution, which is a combination of a multivariate Normal and lognormal distribution. From this distribution, we apply a Bayesian probability framework to derive a non-linear cost function similar to the one that is in current variational data assimilation (DA) applications. Copyright © 2006 Royal Meteorological Society

Keywords: data assimilation; probability; non-Normal; lognormal

Received: 19 December 2005

Revised: 24 March 2006

Accepted: 27 March 2006

1. Introduction

In the field of numerical weather prediction (NWP), there are many different assimilation systems under development and others that are operational, which are derived from Bayesian probability (Lorenc, 1986). These methods are either based on the Kalman filter (Evensen, 1994; Bishop *et al.*, 2001; Houtekamer and Mitchell, 2001, Evensen, 2003), variational data assimilation (VAR) (Parrish and Derber, 1992; Lorenc *et al.*, 2000; Rabier *et al.*, 2000), or a combination of the two techniques (Hamill and Snyder, 2000; Zupanski, 2005; Zupanski and Zupanski, 2006).

In the VAR methods there is a non-linear cost function, which is derived from considering the mode of the analysis probability density function (PDF) (Lorenc, 1986). The mode is the most likely dynamic state and is considered the best fit between the background and observed state. This analysis of PDF is derived from a Bayesian probability problem (Lorenc, 1986). The derivation assumes general PDFs until the errors are defined. At this point, these are assumed to be additive and as such come from a specific class of distributions. The final result is in terms of multivariate Normal distributions.

In recent years, with the introduction of satellites, we see data which is more lognormally rather than Normally distributed (Cohn, 1997; Wang and Sassen, 2002; Sengupta *et al.*, 2004) and it is in Cohn (1997), that a definition for lognormal errors was first presented for NWP. Although it may appear that lognormal variables in NWP is a modern problem, there were data sets in the 1970s that were identified as being lognormally distributed (Mielke *et al.*, 1977).

This then introduces the problem of how to assimilate these variables, given the many different forms of

data assimilation (DA). This problem was addressed by Fletcher and Zupanski (2006) for lognormally distributed observational errors for the variational methods. The result is a non-linear cost function with a Normal background component, which finds the mode of the distribution analysis.

In practice, the assimilation problem is more complex as some of the elements in our state vector or some of the sets of observations may comprise Normal and lognormal variables. For example, in cloud dynamics we have temperature and humidity and for the ocean there is salinity and temperature.

Currently there are three methods to overcome this in VAR DA. The first is to transform the lognormal variable into a Normal variable, assimilate that variable and then transform back to the model space variable. This is often used in 1D retrieval of humidity from radiances (English, 1999; Poli *et al.*, 2002; Deblonde and English, 2003). It is also in use at operational centres, to treat specific humidity in a 3D VAR scheme, Meteorological Service of Canada (Polavarapu *et al.*, 2005). A problem with this is that the mode of the multivariate Normal distribution does not transform back to the mode of the multivariate lognormal distribution.

The second approach is to assimilate the two variables separately. This is possible because of the results in Fletcher and Zupanski (2006). The disadvantage of this is that we are assuming that the variables are uncorrelated. This is commonly assumed for the observations but may not be so justified for the background component. The third method, associated with the observations, is to reject the lognormal variables in a quality control (Lorenc and Hammon, 1988).

An alternative approach, which we show in this article, is to assimilate the two sets of variables simultaneously. This is possible through defining a hybrid PDF, which we do in the next section, and follow the Bayesian framework from (Lorenc, 1986). From this we can define a non-linear cost function that allows the assimilation of Normal and lognormal observations simultaneously (Section 3).

The reason for deriving and proving this hybrid PDF in the next section is that as far as the authors are aware at the time of writing there is no proof of a hybrid distribution of the multivariate Normal/lognormal distribution. Although it may appear trivial to construct this distribution, apparently there are no references to such work in the statistical literature. These hybrid distributions are starting to become more important with the introduction of satellite data into the assimilation methods combined with land observations. Another area where this may become useful is in the assimilation schemes with smaller scale models, i.e. cloud formations and convective modelling, where the common Normal assumption is generally not valid.

Although we may have data that is also not lognormally distributed, it is the first step away from Normal modelling and as such poses problems in itself. If we understand how to assimilate Normal and lognormal variables at the same time, we could move on to possible three way hybrid distributions, for example, Normal–lognormal–Gamma.

The remainder of the article breaks down as follows: Section 2 introduces and proves the multivariate hybrid PDF. Section 3 explains briefly how this hybrid PDF can be used to derive a cost function for variational DA. This article concludes with a plan for further work for this hybrid PDF.

2. Hybrid distribution

In this section, we define and prove a PDF, which is a combination of p Normal variants and q lognormal variants. The PDF is defined in the following theorem.

2.1. Theorem

If p Normal variants and q lognormal variants are given, then the scalar function, $f_{p,q}(x)$, is defined in the following equations:

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

where $N = p + q$, $\tilde{x}^T = (x_p \ln x_q)$, μ is the vector of means and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \cdots & \rho_{1N}\sigma_1\sigma_N \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \cdots & \rho_{2N}\sigma_2\sigma_N \\ \rho_{31}\sigma_3\sigma_1 & \rho_{32}\sigma_3\sigma_2 & \sigma_3^2 & \cdots & \rho_{3N}\sigma_3\sigma_N \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{N1}\sigma_N\sigma_1 & \rho_{N2}\sigma_N\sigma_2 & \rho_{N3}\sigma_N\sigma_3 & \cdots & \sigma_N^2 \end{pmatrix} \quad (2)$$

where $\rho_{i,j}$ $i = 1, 2, \dots, N, j = 1, 2, \dots, N$ are the correlations between the variables, σ_j is the associated standard deviation for \tilde{x}_j , is the covariance matrix, $x_p \in \mathfrak{R}^p$ and $x_q \in \mathfrak{R}^{+q}$, is a multivariate PDF.

2.2. Proof

To prove that Equation (1) defines a PDF, we have two conditions to satisfy. The first is that the function is always positive for all values of \tilde{x} and that the cumulative density function (CDF) integrates to one.

The first point is easy to prove as we can see that the exponential cannot be non-positive, neither can the quotient, the normalizing factor nor the product of the x_i^{-1} s, $i = p + 1, p + 2, \dots, N$.

We now consider the second condition. To prove this we introduce the following change of variable:

$$z_i = \ln x_i \quad (3)$$

for $i = p + 1, p + 2, \dots, N$. The associated limits of integration are $(-\infty, \infty)$ for the transformed variables and finally

$$dy_i = \frac{dx_i}{x_i} \quad (4)$$

An important feature to note here is that the mean vector and the covariance matrix are unchanged. Substituting Equations (3) and (4) combined with the new limits of integration allows us to write the CDF of Equation (1) as

$$CDF = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \times \exp \left\{ -\frac{1}{2} (\hat{x} - \mu)^T \Sigma^{-1} (\hat{x} - \mu) \right\} d\hat{x} \quad (5)$$

where $\hat{x} = (x_p \ z_q)$ is a multivariate Normal random vector and therefore Equation (5) is the CDF of a multivariate Normal distribution (Kotz *et al.*, 2000), which integrates, for all values of \hat{x} , to one and therefore Equation (1) satisfies the second condition and hence it is a multivariate PDF.

Given this new PDF, we now consider how to devise a variational DA scheme in the next section.

3. Application to data assimilation

The important consequence of the hybrid distribution is its application to variational DA. This new distribution allows us to assimilate a set of data that comprises

both Normally and lognormally distributed variables rather than either assimilating the sets of variables separately, where the lognormal components can be assimilated through the method presented in Fletcher and Zupanski (2006), or transform the lognormal component into a Normal variable and use the Normal framework and then transform it back.

To apply this distribution to variational DA we simply follow the argument set out in (Lorenc, 1986), where we define the lognormal errors as in Cohn (1997). For observations these errors are defined as

$$\varepsilon = \frac{y}{h(x)} \tag{6}$$

where y is the vector of observations, h is the non-linear observation operator and the ratio is defined componentwise. This then enables us to write the observational error distribution for a set of mixed data as

$$P_o \propto \left(\prod_{i=p+1}^N \frac{h_i(x)}{y_i} \right) \exp \left\{ -\frac{1}{2} \tilde{\varepsilon}^T R^{-1} \tilde{\varepsilon} \right\} \tag{7}$$

where

$$\tilde{\varepsilon} = \begin{pmatrix} y_p - h_p(x) \\ \ln y_q - \ln h_q(x) \end{pmatrix} \tag{8}$$

Combining with a Normal background framework we obtain the cost function by taking ln of Equation (7). This then results in

$$J(x) = \frac{1}{2} (x - x_b)^T B^{-1} (x - x_b) + \frac{1}{2} \tilde{\varepsilon}^T R^{-1} \tilde{\varepsilon} + \tilde{\varepsilon}^T \begin{pmatrix} 0_p \\ 1_q \end{pmatrix} \tag{9}$$

where x_b is some background state and B is the Normal background covariance matrix.

The associated Jacobian of Equation (9) is given by

$$\frac{\partial J(x)}{\partial x} = B^{-1} (x - x_b) - H^T \frac{\partial \tilde{\varepsilon}}{\partial h} R^{-1} \tilde{\varepsilon} - H^T \frac{\partial \tilde{\varepsilon}}{\partial h} \begin{pmatrix} 0_p \\ 1_q \end{pmatrix} \tag{10}$$

where

$$H = \frac{\partial h}{\partial x}$$

and

$$\frac{\partial \tilde{\varepsilon}}{\partial h} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & \frac{1}{h_{p+1}(x)} & \\ & & & & \ddots \\ & & & & & \frac{1}{h_N(x)} \end{pmatrix}$$

The Hessian associated with the observational part of Equation (11) is given by

$$\begin{aligned} \frac{\partial^2 J(x)}{\partial x_i \partial x_j} = & \left[H^T \frac{\partial \tilde{\varepsilon}}{\partial h} R^{-1} \frac{\partial \tilde{\varepsilon}}{\partial h} H + H^T \frac{\partial \tilde{\varepsilon}}{\partial h} \begin{pmatrix} 0_{pp} & 0_{pq} \\ 0_{qp} & I_q \end{pmatrix} \right. \\ & \times \left. \frac{\partial \tilde{\varepsilon}}{\partial h} H \right]_{i,j} + \left[G_i^T \begin{pmatrix} I_p & 0_{pq} \\ 0_{qp} & \frac{\partial^2 \hat{\varepsilon}}{\partial h^2} \end{pmatrix} R^{-1} \tilde{\varepsilon} \right. \\ & \left. + G_i^T \begin{pmatrix} I_p & 0_{pq} \\ 0_{qp} & \frac{\partial^2 \hat{\varepsilon}}{\partial h^2} \end{pmatrix} \begin{pmatrix} 0_p \\ 1_q \end{pmatrix} \right]_j \tag{11} \end{aligned}$$

where

$$G_i = \frac{\partial}{\partial x_i} \left(\frac{\partial h}{\partial x} \right)$$

and the hat represents the lognormal part of the observation error.

An important feature of the derivatives above is that they have a structure similar to their counterparts from a multivariate Normal framework. A consequence of this is the ability to apply a Hessian pre-conditioner (Zupanski, 2005), which aids in the minimization of the cost function (Axelsson and Barker, 1984). The pre-conditioner for the lognormal observations is derived in Fletcher and Zupanski (2006).

The point that we are making here is that we no longer have to ignore or transform non-Normal variables into their Normal counterpart. We therefore reduce the error associated with the transform but also remove the non-uniqueness associated with the multivariate lognormal median, which is the statistics that is found through transforming the lognormal variable and assimilating through the Normal framework (Fletcher and Zupanski, 2006).

4. Conclusions and further work

In this article, we have defined a scalar function and have proved that it is a hybrid probability density function, which is a combination of p Normal and q lognormal variants. Following the Bayesian framework from Lorenc (1986) combined with the lognormal error definition from Cohn (1997) and the lognormal cost function from Fletcher and Zupanski (2006) we are able to define a framework for data sets of mixed types. We have derived the Jacobian and the Hessian of the hybrid cost function, which are useful in the minimization of the cost function. These can be used without too much modification to the current Normal framework because of this distribution being unimodal.

The reason we derive this hybrid distribution is to illustrate that we can define frameworks for variational and ensemble methods used in weather and ocean prediction that allows us to assimilate variables, which are from different PDFs simultaneously. The impact would allow us to assimilate satellite data at the same

time as land observations, which are conventionally Normally distributed.

The plan for this work is to implement this new cost function with the maximum likelihood ensemble filter (MLEF) (Zupanski, 2005), which is an ensemble filter combined with a cost function similar to that from 3D VAR. We plan to use this method with Rossby-Haurwitz waves in a 2D spherical shallow water equations model that generates flows similar to that of the full atmosphere (Daley, 1996).

As mentioned in the introduction this is the first step in allowing us to use observations, which may not be Normally distributed. This framework allows us to assimilate Normal and lognormal variables simultaneously but more work is needed to identify the distribution of the variables.

Therefore an extension to this work is to classify data sets into their distribution and then define the error either as an additive or multiplicative. A consequence of this aids the ensemble methods, which use ensemble means to generate their error statistics (Bishop *et al.*, 2001). If an ensemble method has lognormal variables then a mean can still be calculated by calculating the product of the ensembles and then taking the number of the ensemble root. The problem is whether this statistic is the best to represent that distribution (Fletcher and Zupanski, 2006).

Acknowledgements

We are grateful for many useful discussions with Drs Laura Fowler, Manajit Sengupta, Tomislava Vukicevic and Dusanka Zupanski about the motivation for the need for this research. We are also thankful for the useful comments from the two anonymous reviewers of this manuscript. We would also like to thank the National Center for Atmospheric Research for the use of the super computing facilities. This research was supported by the National Science Foundation Collaboration in Mathematical Geosciences (grant No. 0327651).

References

- Axelsson O, Barker VA. 1984. *Finite Element Solution of Boundary Value Problems. Theory and Computation*. Academic Press: Orlando.
- Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review* **129**: 420–436.
- Cohn SE. 1997. An introduction to estimation theory. *Journal of the Meteorological Society of Japan* **75**: 257–288.
- Daley R. 1996. *Atmospheric Data Analysis*. Cambridge University Press: Cambridge.
- Deblonde G, English S. 2003. One-dimensional variational retrievals from SSMIS-Simulated observations. *Journal of Applied Meteorology* **42**: 1406–1420.
- English SJ. 1999. Estimation of temperature and humidity profile information from microwave radiances over different surface types. *Journal of Applied Meteorology* **38**: 1526–1541.
- Evensen G. 1994. Data assimilation with a non-linear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research* **99**(C5): 10,143–10,162.
- Evensen G. 2003. The ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics* **53**: 343–367.
- Fletcher SJ, Zupanski M. 2006. Accepted: A data assimilation method for lognormally distributed observational errors. *Quarterly Journal of the Royal Meteorological Society* (in press).
- Hamill TM, Snyder C. 2000. A hybrid ensemble Kalman filter-3d variational analysis system. *Monthly Weather Review* **128**: 2905–2919.
- Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review* **129**: 123–137.
- Kotz S, Balakrishnan N, Johnson NL. 2000. *Continuous Multivariate Distributions, Volume 1: Models and Applications* 2nd edn. John Wiley and Sons: New York.
- Lorenz AC. 1986. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **112**: 1177–1194.
- Lorenz AC, Hammon O. 1988. Effective quality control of observations using Bayesian methods. Theory, and a practical implementation. *Quarterly Journal of the Royal Meteorological Society* **114**: 515–543.
- Lorenz AC, Ballard SP, Bell RS, Ingleby NB, Andrews PLF, Barker DM, Bray JR, Clayton AM, Dalby T, Li D, Payne TJ, Saunders FW. 2000. The Met. Office global three dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society* **126**: 2991–3012.
- Mielke PW Jr, Williams JS, Wu S-C. 1977. Covariance analysis techniques based upon bivariate log-Normal distribution with weather modification application. *Journal of Applied Meteorology* **16**: 183–187.
- Parrish DF, Derber JC. 1992. The National Meteorological Center's Spectral Statistical-Interpolation Analysis System. *Monthly Weather Review* **120**: 1747–1763.
- Polavarapu S, Ren S, Rochon Y, Sankey D, Ek N, Koshyk J, Tarasick D. 2005. Data assimilation with the Canadian middle atmosphere model. *Atmosphere-Ocean* **43**(1): 77–100.
- Poli P, Joiner J, Kursinski ER. 2002. 1DVAR analysis of temperature and humidity using GPS radio occultation refractivity data. *Journal of Geophysical Research* **107**(D20): 4448–4468.
- Rabier F, Jarvinen H, Klinker E, Mahouf J-F, Simmons A. 2000. The ECMWF implementation operational of four dimensional variational assimilation. Part I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society* **126A**: 1143–1170.
- Sengupta M, Clothiaux EE, Ackerman TP. 2004. Climatology of warm boundary layer clouds at the ARM SCP site and their comparison to models. *Journal of Climate* **17**: 4760–4782.
- Wang Z, Sassen K. 2002. Cirrus cloud microphysics property retrieval using lidar and radar measurements. Part II: Midlatitude cirrus microphysical and radiative properties. *Journal of Atmospheric Sciences* **59**: 2291–2302.
- Zupanski M. 2005. Maximum Likelihood Ensemble Filter. Part I: Theoretical aspects. *Monthly Weather Review* **133**: 1710–1726.
- Zupanski D, Zupanski M. 2006. Model error estimation employing ensemble data assimilation approach. *Monthly Weather Review* **134**: 1337–1354.