École Doctorale des Sciences de l'Environnement d'Île-de-France

Année Universitaire 2011-2012

Modélisation Numérique de l'Écoulement Atmosphérique et Assimilation de Données

Olivier Talagrand

Cours 4

5 Avril 2013

$z_1 = x + \zeta_1$ density function $p_1(\zeta) \propto \exp[- (\zeta^2)/2s_1]$

$z_2 = x + \zeta_2$ density function $p_2(\zeta) \propto \exp[- (\zeta^2)/2s_2]$

ζ_1 and ζ_2 mutually independent

Conditional probability $P(x = \xi \mid z_1, z_2)$?

$z_1 = x + \zeta_1$ density function $p_1(\zeta) \propto \exp[-(\zeta^2)/2s_1]$

$z_2 = x + \zeta_2$ density function $p_2(\zeta) \propto \exp[-(\zeta^2)/2s_2]$

ζ_1 and ζ_2 mutually independent

$x = \xi \;\Leftrightarrow\; \zeta_1 = z_1 - \xi$ and $\zeta_2 = z_2 - \xi$

- $P(x = \xi \mid z_1, z_2) \propto p_1(z_1 - \xi)\, p_2(z_2 - \xi)$

$$\propto \exp[-(\xi - x^a)^2/2p^a]$$

where $1/p^a = 1/s_1 + 1/s_2$, $x^a = p^a\,(z_1/s_1 + z_2/s_2)$

Conditional probability distribution of x, given z_1 and z_2 : $\mathcal{N}[x^a, p^a]$

$p^a < (s_1, s_2)$ independent of z_1 and z_2

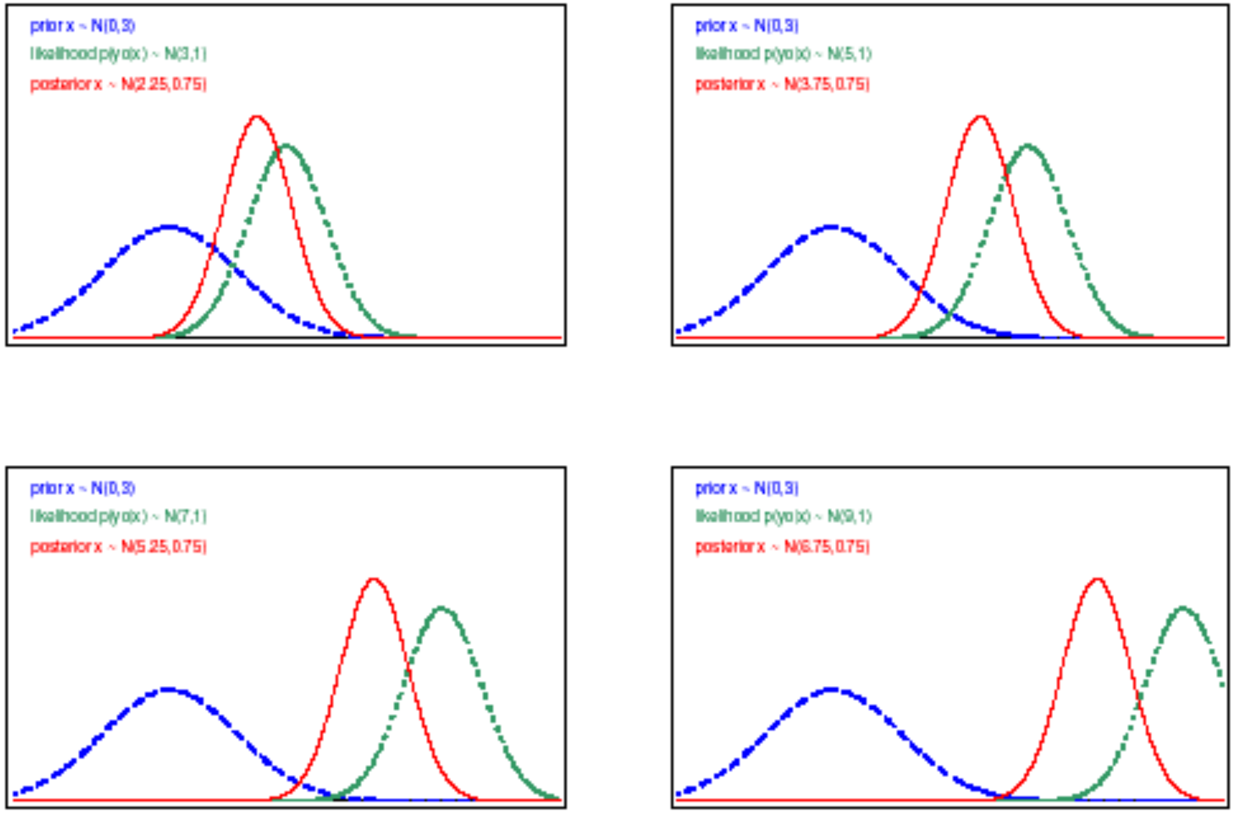


Fig. 1.1: Prior pdf $p(x)$ (dashed line), posterior pdf $p(x|y^o)$ (solid line), and Gaussian likelihood of observation $p(y^o|x)$ (dotted line), plotted against x for various values of y^o. (Adapted from Lorenc and Hammon 1988.)

Conditional expectation x^a minimizes following scalar *objective function*, defined on ξ-space

$$\xi \rightarrow \mathcal{J}(\xi) \equiv (1/2)\,[(z_1 - \xi)^2 / s_1 + [(z_2 - \xi)^2 / s_2\,]$$

In addition

$$p^a = 1/\,\mathcal{J}''(\xi)$$

Conditional probability distribution in Gaussian case

$$P(x = \xi \mid z_1, z_2) \propto \exp[\, - \underbrace{(\xi - x^a)^2/2p^a}_{\mathcal{J}(\xi)}]$$

$z_1 = x + \zeta_1$
$z_2 = x + \zeta_2$

Same as before, but ζ_1 and ζ_2 are now distributed according to exponential law with parameter a, *i. e.*

$$p(\zeta) \propto \exp[-|\zeta|/a] \quad ; \quad \mathrm{Var}(\zeta) = 2a^2$$

Conditional probability density function is now uniform over interval $[z_1, z_2]$, exponential with parameter $a/2$ outside that interval

$$E(x \mid z_1, z_2) = (z_1+z_2)/2$$

$\mathrm{Var}(x \mid z_1, z_2) = a^2\,(2\delta^3/3 + \delta^2 + \delta + 1/2)\,/\,(1 + 2\delta)$, with $\delta = |z_1 - z_2|/(2a)$
Increases from $a^2/2$ to ∞ as δ increases from 0 to ∞. Can be larger than variance $2a^2$ of original errors (probability 0.08)

(Entropy $-\int p \ln p$ always decreases in bayesian estimation)

Bayesian estimation

State vector x, belonging to *state space* $\mathcal{S}$ ($\dim \mathcal{S} = n$), to be estimated.

Data vector z, belonging to *data space* $\mathcal{D}$ ($\dim \mathcal{D} = m$), available.

$$z = F(x, \zeta) \qquad (1)$$

where ζ is a random element representing the uncertainty on the data (or, more precisely, on the link between the data and the unknown state vector).

For example

$$z = \Gamma x + \zeta$$

Bayesian estimation (continued)

Probability that $x = \xi$ for given ξ ?

$$x = \xi \quad \Rightarrow \quad z = F(\xi, \zeta)$$

$$P(x = \xi \mid z) = P[z = F(\xi, \zeta)] \,/ \int_{\xi'} P[z = F(\xi', \zeta)]$$

Unambiguously defined iff, for any ζ, there is at most one x such that (1) is verified.

⇔ data contain information, either directly or indirectly, on any component of x. *Determinacy* condition.

Bayesian estimation is however impossible in its general theoretical form in meteorological or oceanographical practice because

- It is impossible to explicitly describe a probability distribution in a space with dimension even as low as $n \approx 10^3$, not to speak of the dimension $n \approx 10^{6-9}$ of present Numerical Weather Prediction models.

- Probability distribution of errors on data very poorly known (model errors in particular).

One has to restrict oneself to a much more modest goal. Two approaches exist at present

- Obtain some 'central' estimate of the conditional probability distribution (expectation, mode, …), plus some estimate of the corresponding spread (standard deviations and a number of correlations).

- Produce an ensemble of estimates which are meant to sample the conditional probability distribution (dimension $N \approx O(10\text{-}100)$).

Random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T = (x_i)$ (*e. g.* pressure, temperature, abundance of given chemical compound at n grid-points of a numerical model)

- Expectation $E(\boldsymbol{x}) \equiv [E(x_i)]$; centred vector $\boldsymbol{x}' \equiv \boldsymbol{x} - E(\boldsymbol{x})$

- Covariance matrix

$$E(\boldsymbol{x}'\boldsymbol{x}'^T) = [E(x_i' x_j')]$$

  dimension nxn, symmetric non-negative (strictly definite positive except if linear relationship holds between the x_i''s with probability 1).

- Two random vectors
  $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$
  $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$

$$E(\boldsymbol{x}'\boldsymbol{y}'^T) = E(x_i' y_j')$$

  dimension nxp

Random function $\Phi(\xi)$ (field of pressure, temperature, abundance of given chemical compound, … ; ξ is now spatial and/or temporal coordinate)

- Expectation $E[\Phi(\xi)]$; $\Phi'(\xi) \equiv \Phi(\xi) - E[\Phi(\xi)]$
- Variance $Var[\varphi(\xi)] = E\{[\varphi'(\xi)]^2\}$

- Covariance function

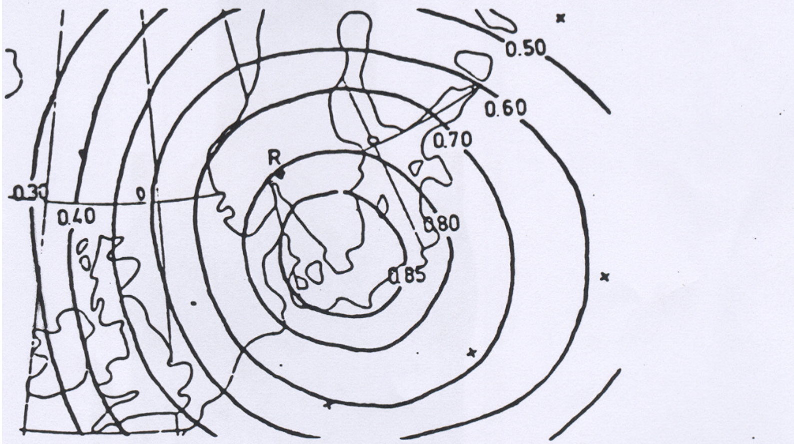
$$(\xi_1, \xi_2) \rightarrow C_\Phi(\xi_1, \xi_2) \equiv E[\Phi'(\xi_1)\ \Phi'(\xi_2)]$$

- Correlation function

$$Cor_\varphi(\xi_1, \xi_2) \equiv E[\Phi'(\xi_1)\ \Phi'(\xi_2)] / \{Var[\Phi(\xi_1)]\ Var[\Phi(\xi_2)]\}^{1/2}$$

•

.: Isolines for the auto-correlations of the 500 mb geopotential between the station in Hannover and surrounding stations.
From Bertoni and Lund (1963)



: Isolines of the cross-correlation between the 500 mb geopotential in station 01 384 (R) and the surface pressure in surrounding stations.

After N. Gustafsson

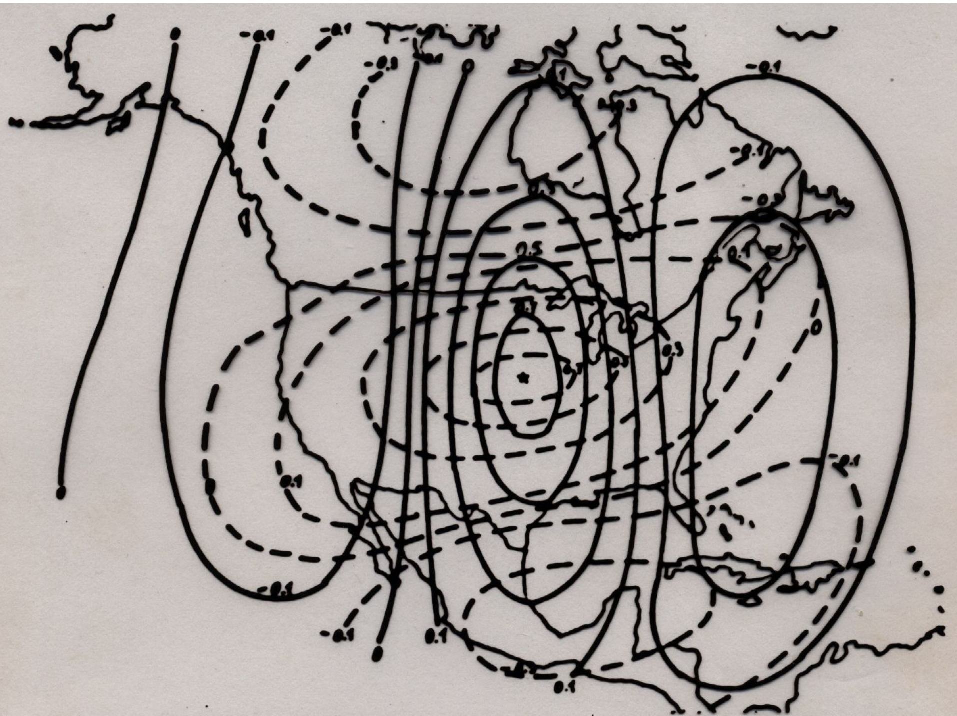


Figure 4.2.4.3: Isolines for the auto-correlation of the 500 mb u-wind component (dashed line) and the auto-correlation of the 500 mb v-wind component (full line). The "star" indicates the position of the reference station. (From Buel (1972).

After N. Gustafsson

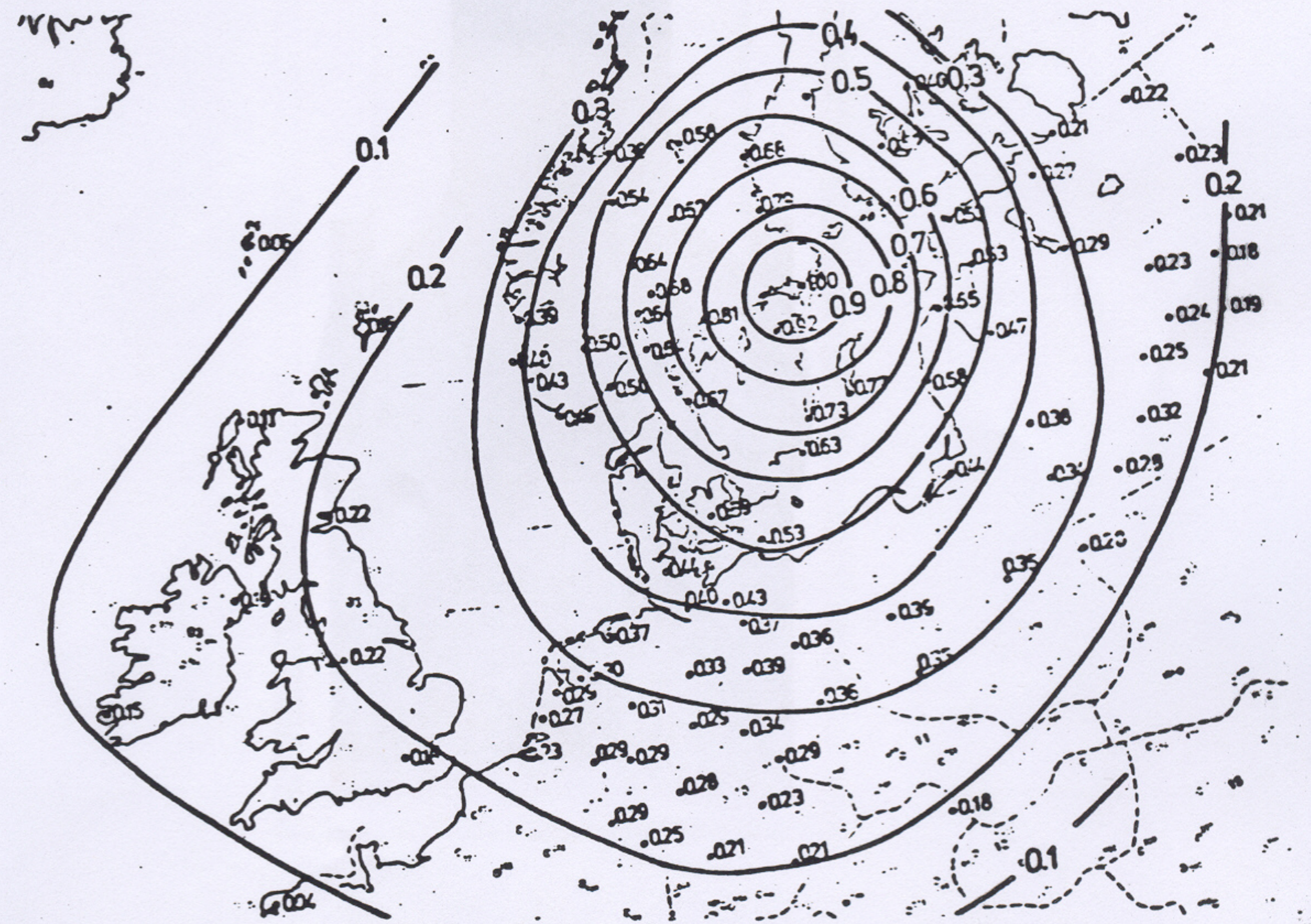


Figure 5.1.1.4.1 Auto-correlation of errors in 12h numerical forecasts of surface pressure in a reference station (Stockholm) and other stations.

After N. Gustafsson

Optimal Interpolation

Random field $\Phi(\xi)$

Observation network $\xi_1, \xi_2, \ldots, \xi_p$

For one particular realization of the field, observations

$$y_j = \Phi(\xi_j) + \varepsilon_j \quad , j = 1, \ldots, p \qquad , \qquad \text{making up vector } \boldsymbol{y} = (y_j)$$

Estimate $x = \Phi(\xi)$ at given point ξ, in the form

$$x^a = \alpha + \Sigma_j \beta_j y_j = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{y} \qquad , \qquad \text{where } \boldsymbol{\beta} = (\beta_j)$$

α and the β_j's being determined so as to minimize the expected quadratic estimation error $E[(x\text{-}x^a)^2]$