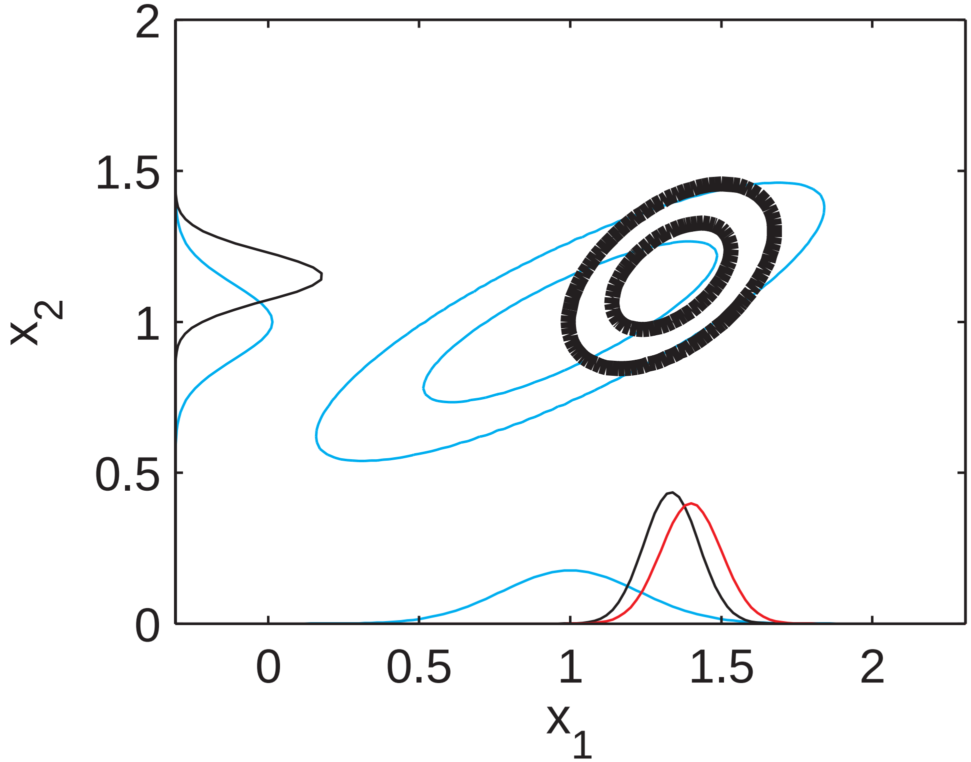
Basics of Data Assimilation



- ▷ Chris Snyder (NCAR)

Why Data Assimilation?

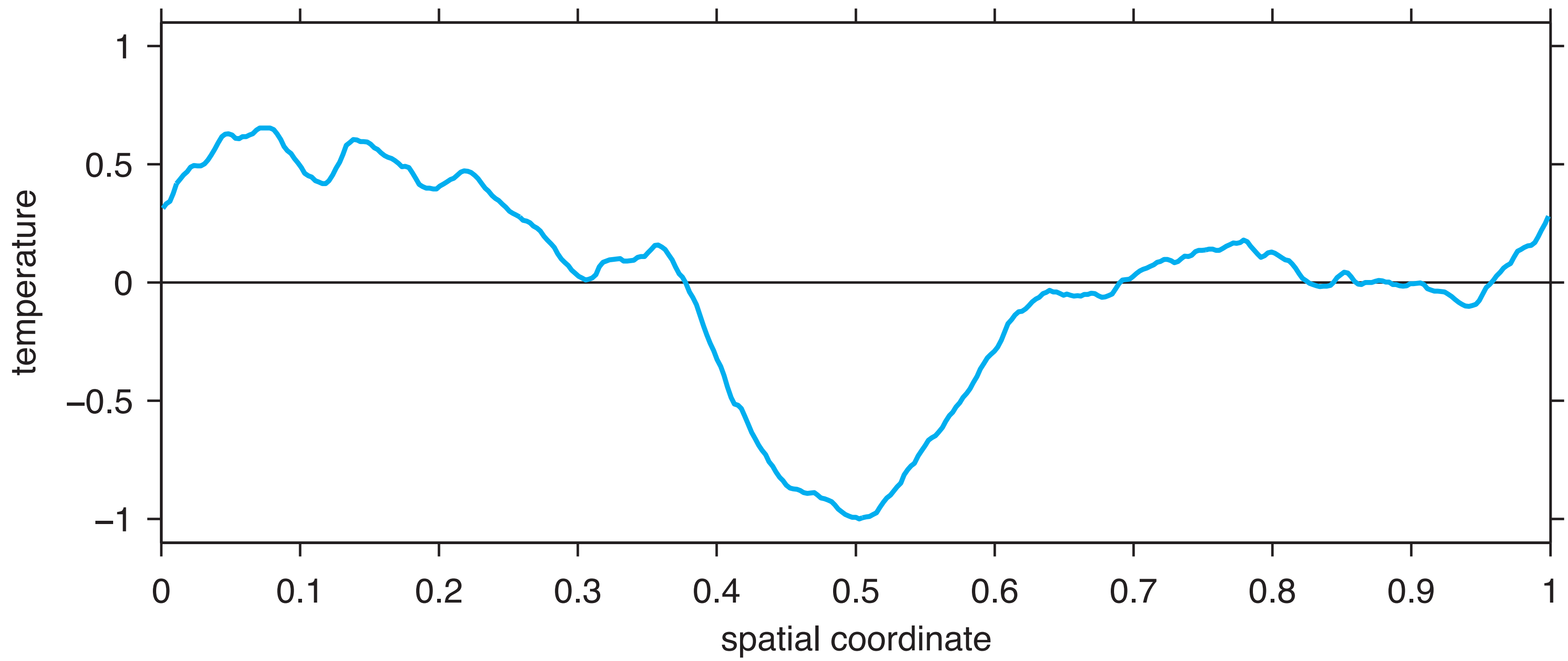
We need initial conditions for NWP

We have limited observations

Simplest: interpolate between observations

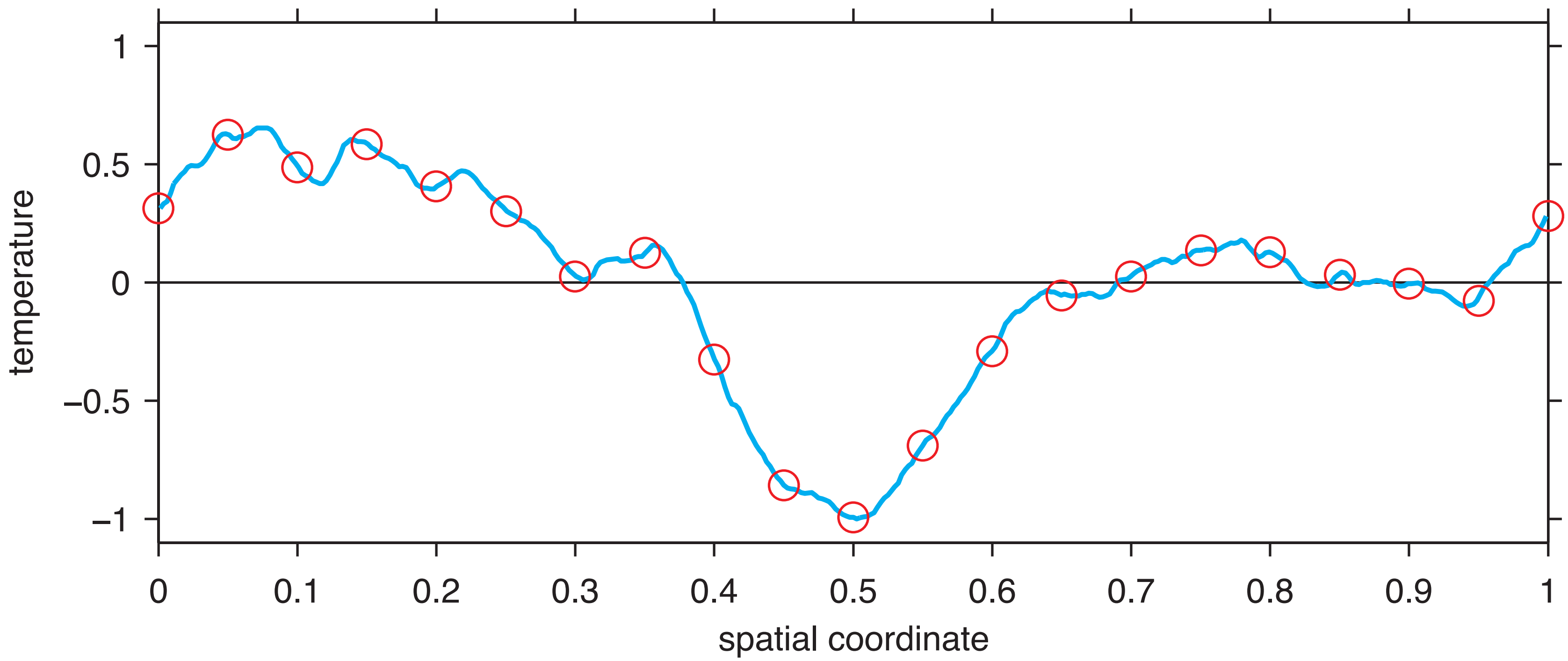
- ▷ *objective analysis*

Function Fitting



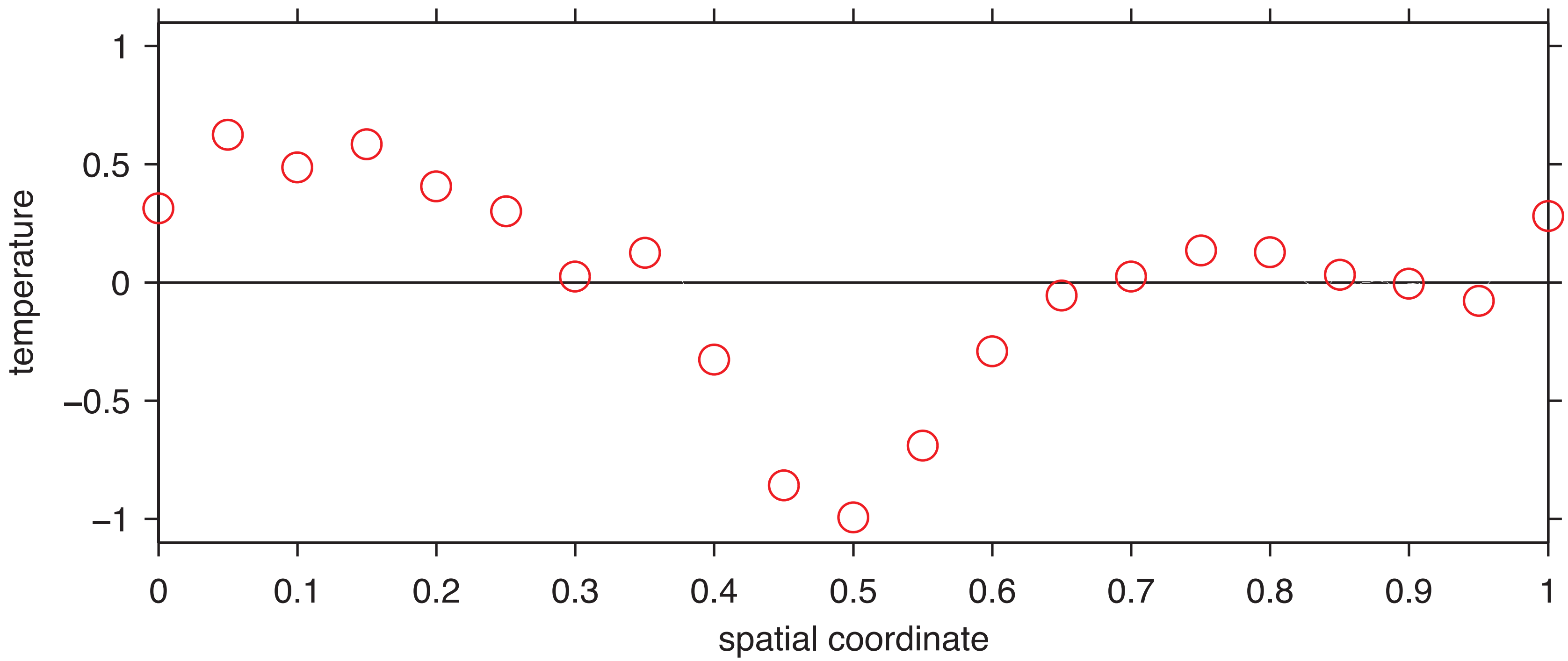
- field to be estimated

Function Fitting



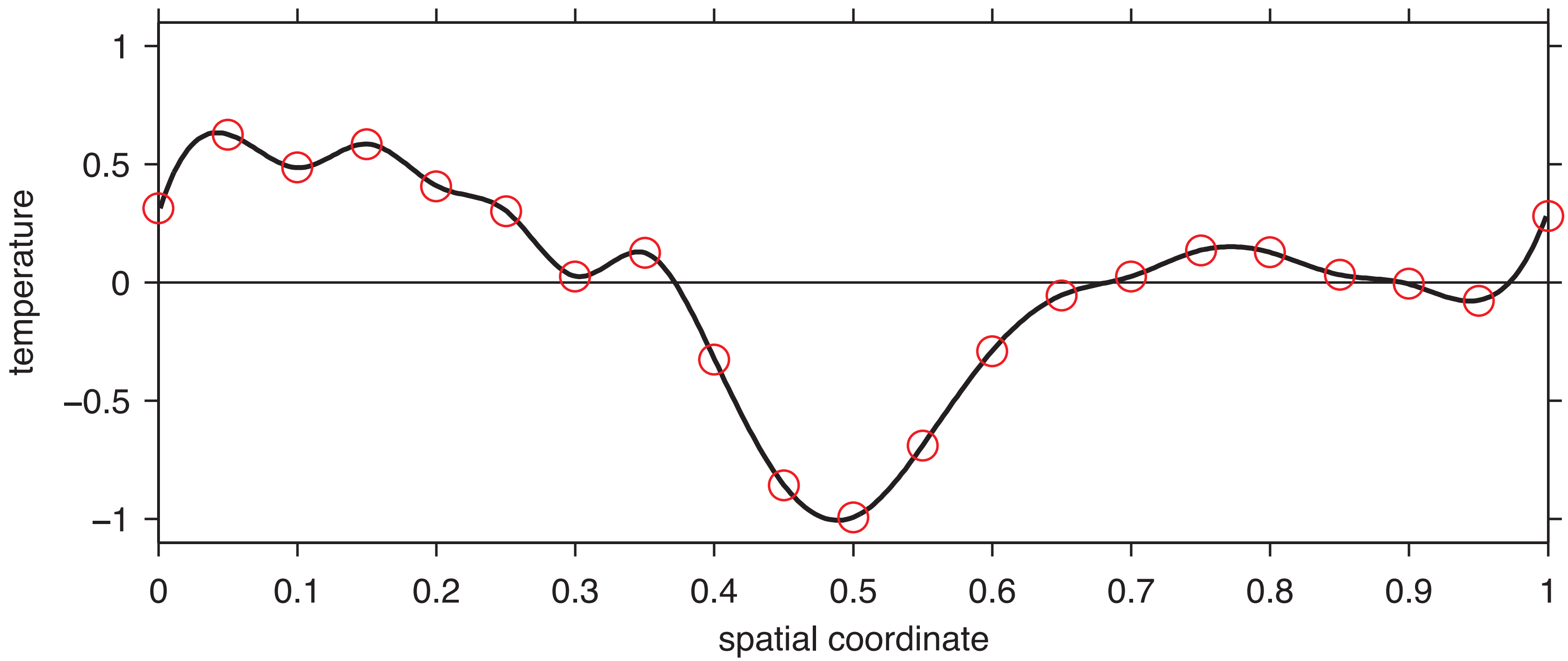
- ▷ observations at discrete points; no obs error

Function Fitting



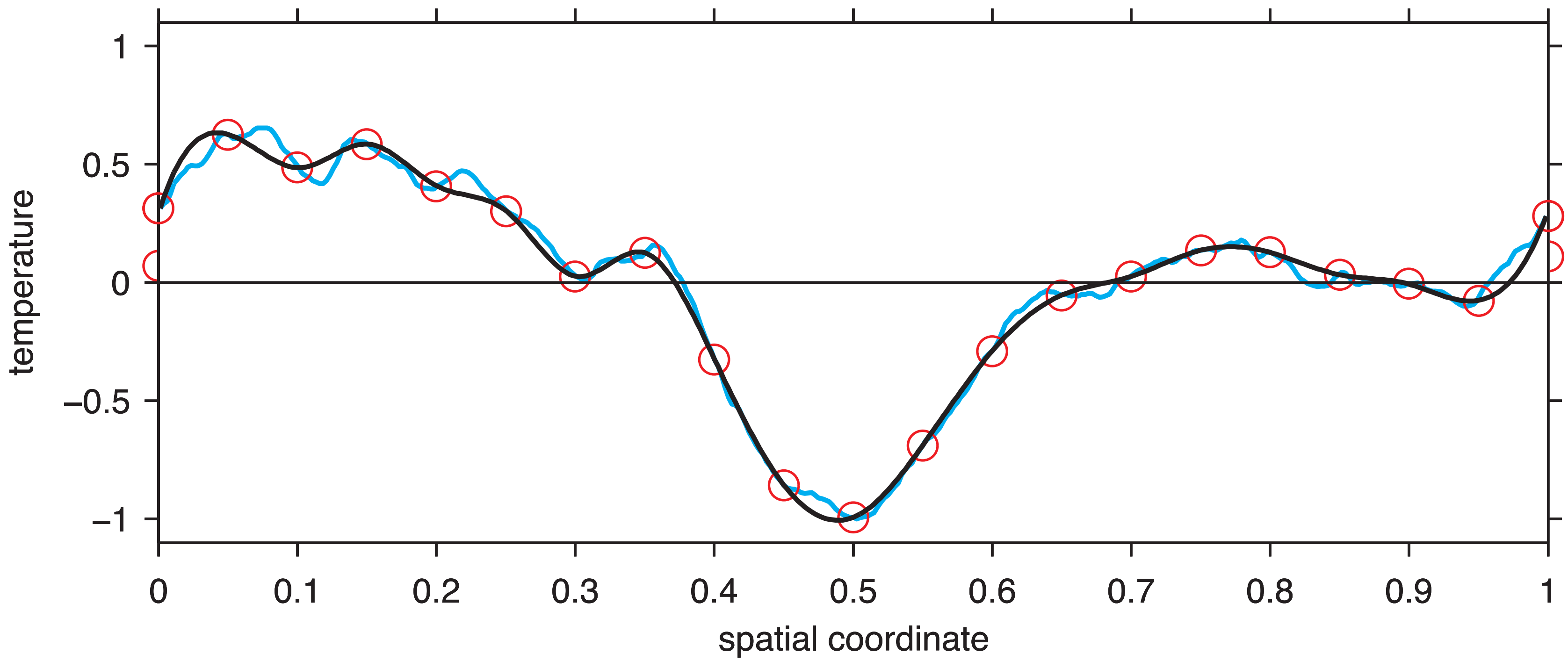
- observations at discrete points; no obs error

Function Fitting



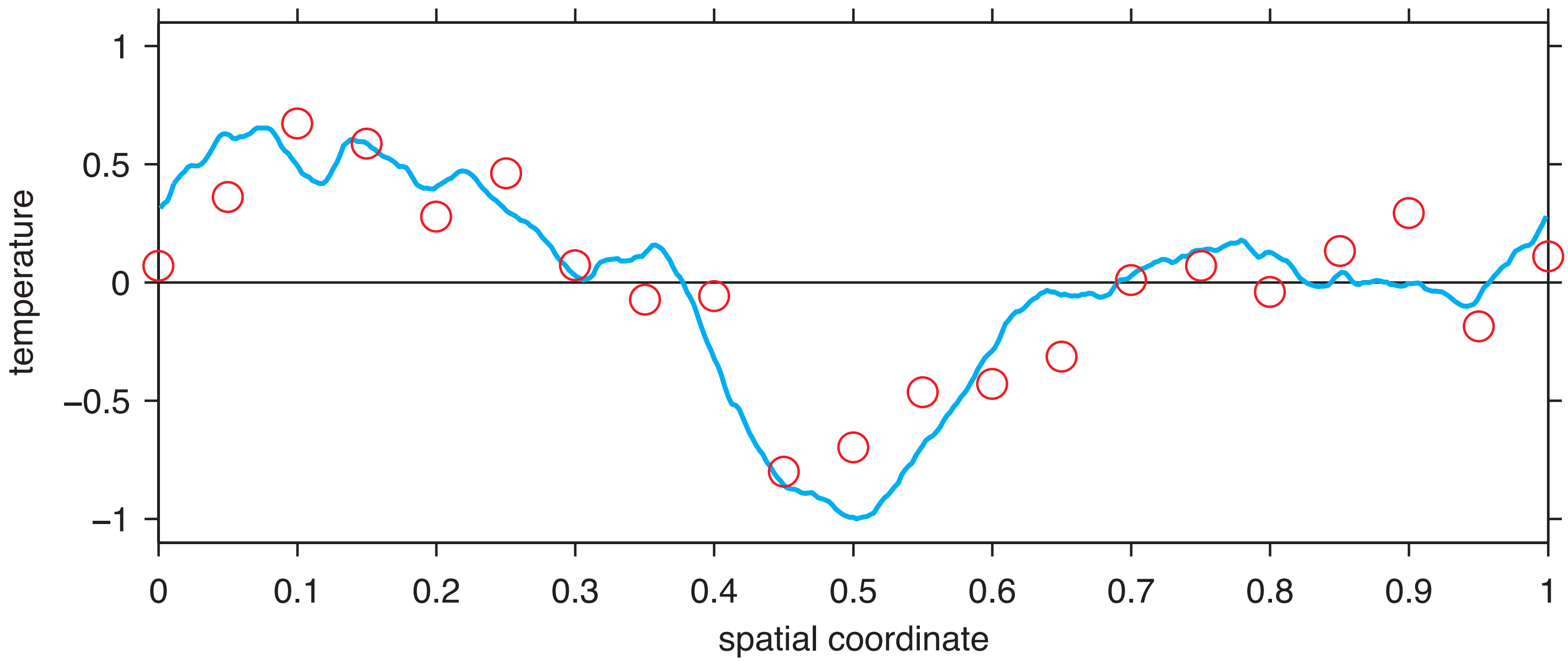
- ▷ cubic-spline interpolation

Function Fitting



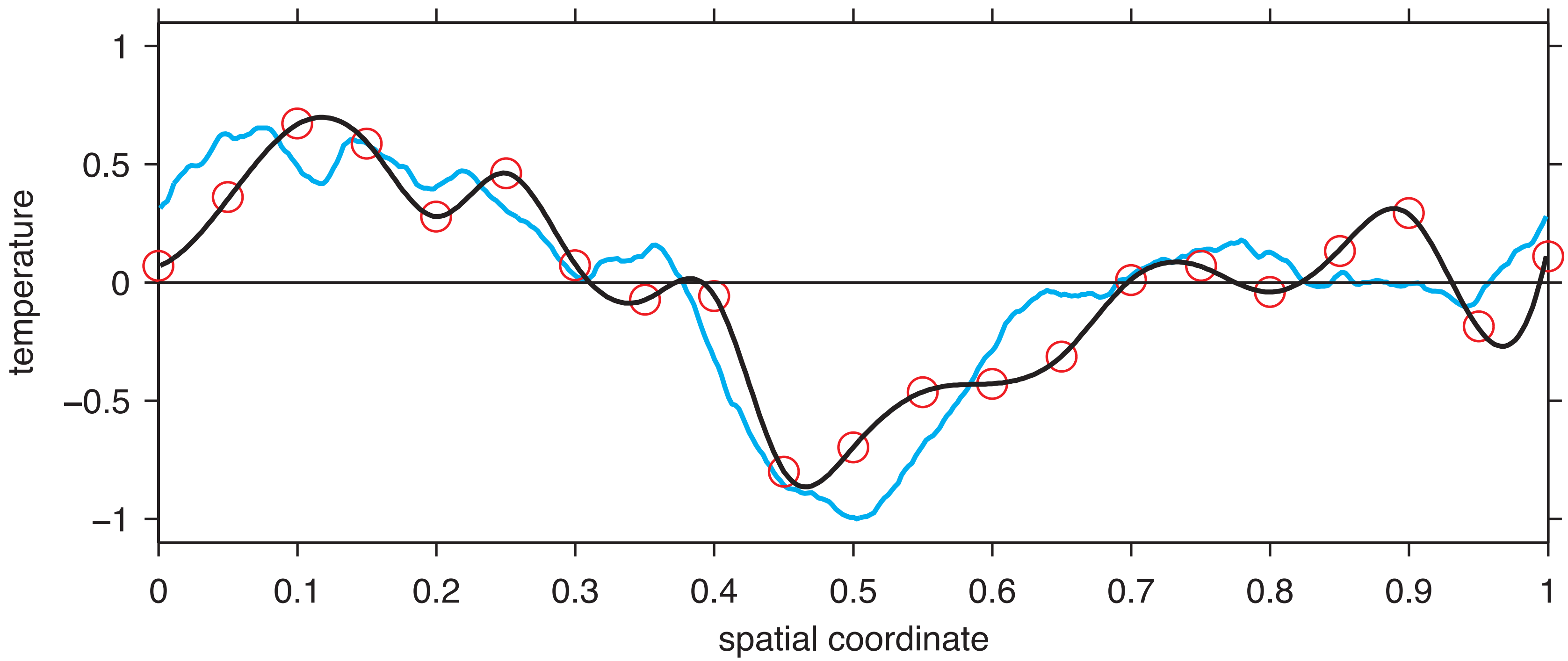
- interpolation gives good estimate of scales resolved by obs

Function Fitting



▷ imperfect observations (std. dev. 0.2)

Function Fitting



- ▷ small scales lost in observational noise

Why Data Assimilation? (II)

Real observing networks

- ▷ observations are imperfect
- ▷ obs may be *indirect:*
  measured variables differ from forecast-model variables
- ▷ inhomogeneous in space and time $\rightarrow$ gaps

Use other information to fill gaps

- ▷ e.g., previous forecast (*background* or *first guess*)
- ▷ How to combine with observations?

Observations (and forecasts!) have errors

- ▷ How to account for this?

Indirect observations

- ▷ How to spread information from observed variables to model variables?

A Basic Example

Two instruments, both measure a quantity x

- observations have form
$$\begin{aligned} y_1 &= x + \epsilon_1, \\ y_2 &= x + \epsilon_2 \end{aligned}$$
- observation errors ϵ_1, ϵ_2 are random but we know something about their statistics

Given observations y_1 and y_2, estimate x

- since estimate is imperfect, want estimate of its accuracy as well.

The Minimum Variance Solution

Acknowledge probabilistic aspects of problem

- obs errors are random
- can't know true value of x; can only evaluate expected errors

Seek unbiased x_a with minimum expected squared error

- $x_a - x$ is error
- require
  $$E(x_a - x) = 0$$
- minimize
  $$E\left((x_a - x)^2\right) = \mathsf{Var}\left(x_a - x\right)$$

The Minimum Variance Solution (cont.)

Spse x_a is linear combination of observations,

$$x_a = \alpha_1 y_1 + \alpha_2 y_2 = (\alpha_1 + \alpha_2)x + \alpha_1 \epsilon_1 + \alpha_2 \epsilon_2$$

Then,

$$E(x_a - x) = 0 \quad \Rightarrow \quad \alpha_1 + \alpha_2 = 1 \quad \text{if } E(\epsilon_i) = 0$$

Next,

$$\begin{aligned} E\left((x_a - x)^2\right) &= E\left[\{\alpha_1(y_1 - x) + \alpha_2(y_2 - x)\}^2\right] \\ &= E\left[\alpha_1^2\epsilon_1^2 + \alpha_2^2\epsilon_2^2\right] \qquad \text{if } E(\epsilon_1\epsilon_2) = 0 \\ &= \alpha_1^2\sigma_1^2 + (1-\alpha_1)^2\sigma_2^2, \end{aligned}$$

where $\sigma_i^2 = \mathrm{Var}\,(\epsilon_i)$

The Minimum Variance Solution (cont.)

Minimizing w.r.t. α's then gives

$$\alpha_1 = \sigma_2^2/(\sigma_1^2 + \sigma_2^2), \quad \alpha_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$$

Properties

- solution depends only on observation-error variances
- observations with large errors receive small weight
- expected squared error
  $$E\left((x_a - x)^2\right) = \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 = \sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$$
- error of x_a is (on average) smaller than error of either observation

Best linear unbiased estimator (BLUE)

The Bayesian Solution

True state is unknown

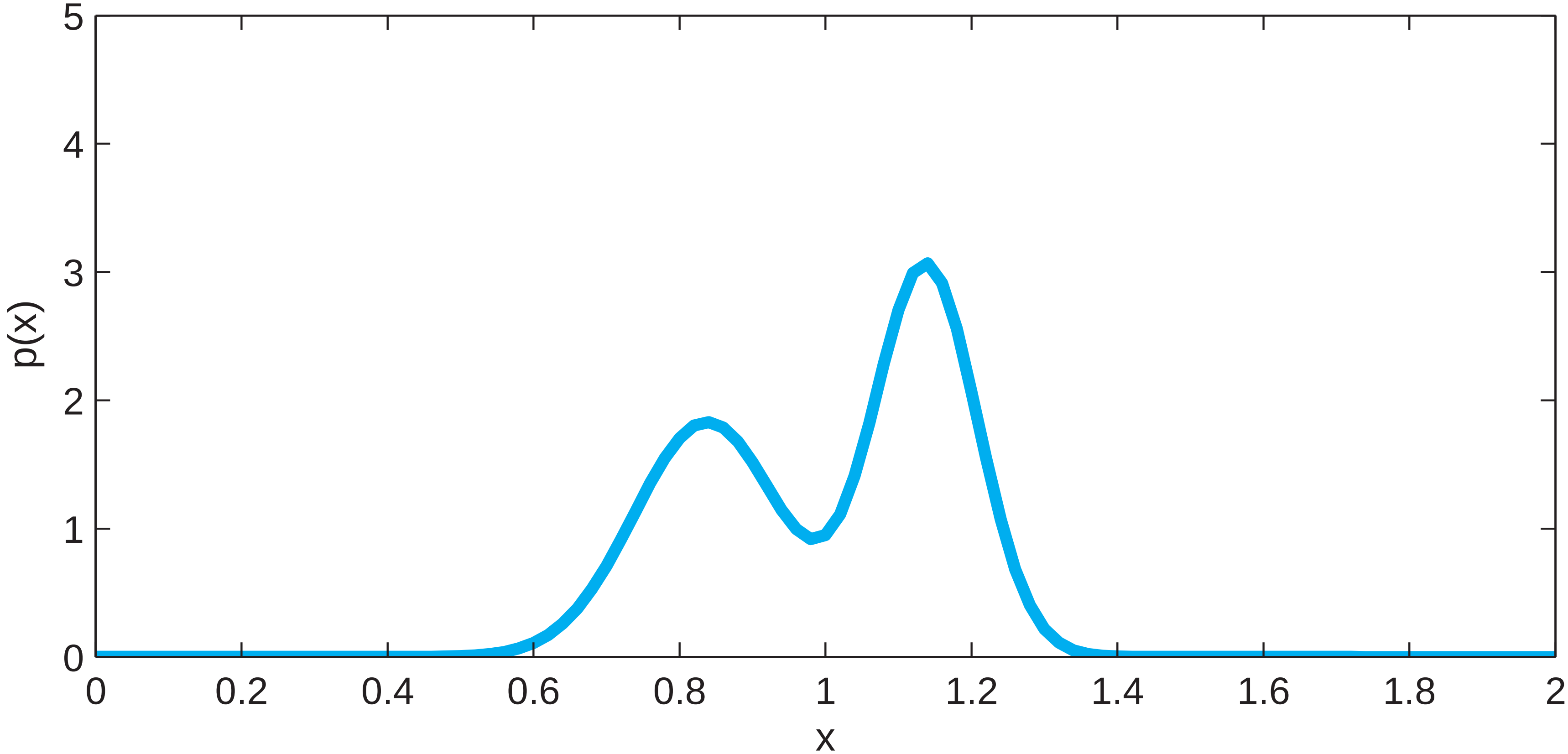
- observations, models both have random errors
- wish to calculate $p(x|y_1, y_2)$, probability density of x given y_1 and y_2. Also known as the *posterior* density.

Bayes rule

- $p(x|y_1, y_2) \propto p(y_2|x, y_1)p(x|y_1)$
- if obs errors independent, $p(y_2|x, y_1) = p(y_2|x)$

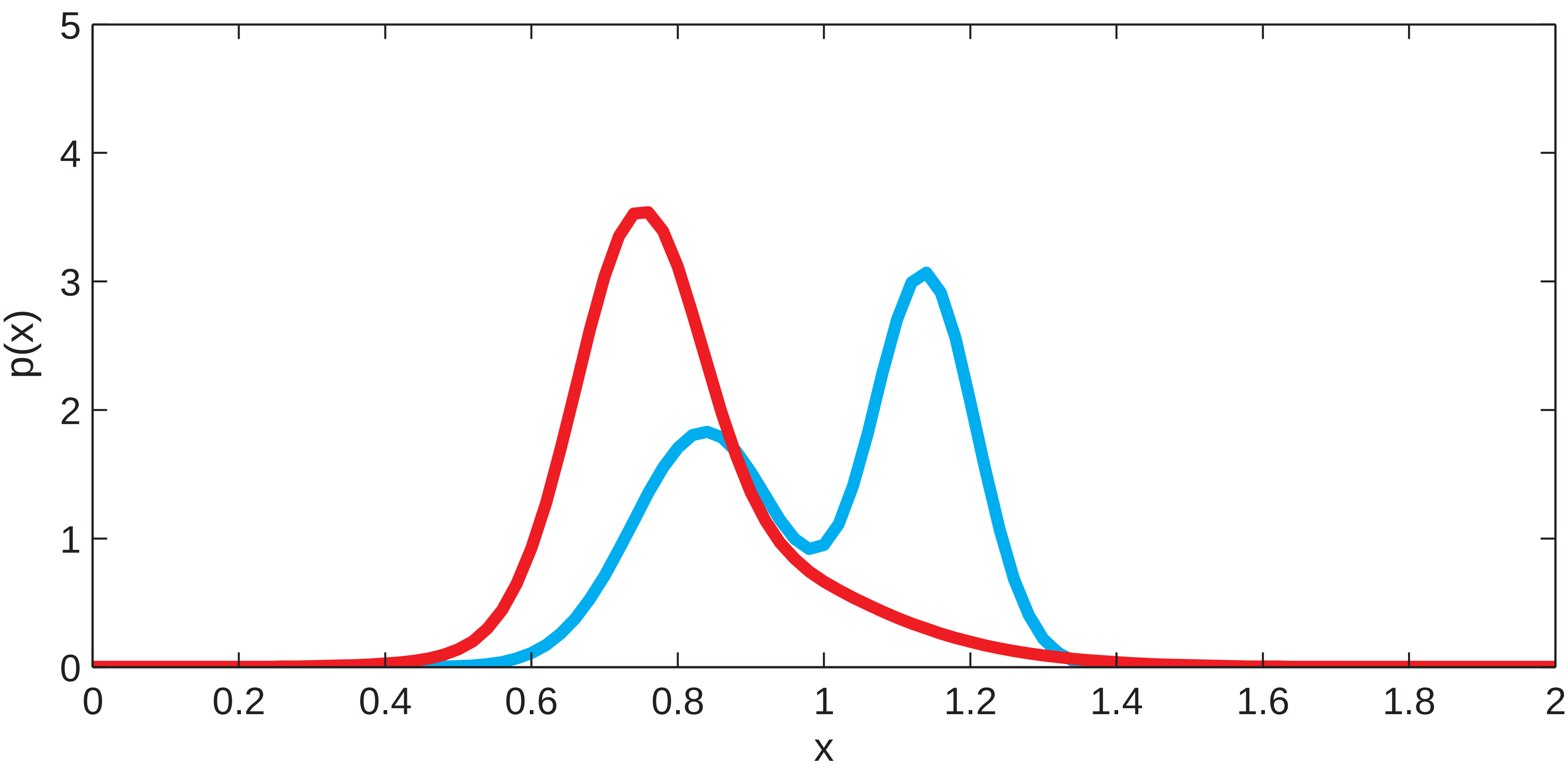
Bayes Illustrated

- $p(x|y_1)$ (blue)



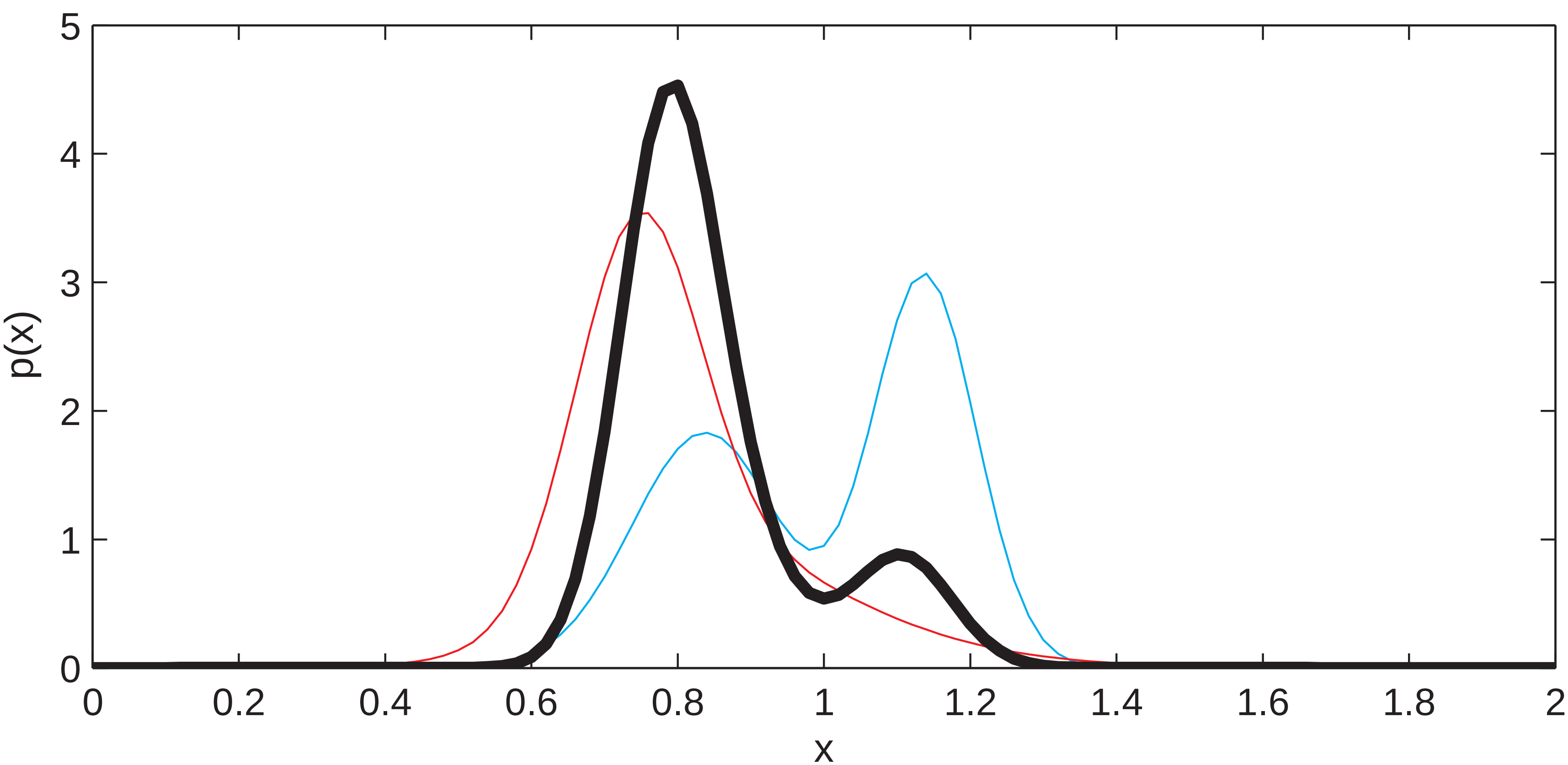
Bayes Illustrated (cont.)

- $p(x|y_1)$ (blue)
- $p(y|x)$ for $y = 0.75$ (red)



Bayes Illustrated (cont.)

- $p(x|y_1)$ (blue)
- $p(y|x)$ for $y = 0.75$ (red)
- $p(x|y) \propto p(y|x)p(x)$ (black)



Bayes Solution for Gaussian Errors

Suppose densities are Gaussian

- e.g., of the form $p(x) = c\exp(-\frac{1}{2}(x-\bar{x})^2/\sigma^2)$

Bayes rule involves multiplication of densities

- products of Gaussians amount to adding exponents
- for two observations, posterior density proportional to

$$\exp[-\tfrac{1}{2}J(x)] = \exp[-\tfrac{1}{2}(x-y_1)^2/\sigma_1^2 - \tfrac{1}{2}(x-y_2)^2/\sigma_2^2]$$

Yields same answer as minimum-variance approach

- left as exercise
- hint: write $\exp[-\frac{1}{2}J(x)] = \exp[-\frac{1}{2}(x-x_a)^2/\sigma_a^2]$

Background Forecasts

Often, have prior information in addition to observations

- climatology
- forecast from earlier time (first guess, background)

In NWP, background forecasts often as accurate as observations

- propagates information from previous obs forward in time

Need to estimate error statistics for DA

Many variables

Notation

- $\mathbf{x}$ = continuous state of system *projected* onto discrete basis, e.g. grid-point values or Fourier coefficients
- $\mathbf{y}$ = all observations concatenated into single vector

Observation Operator and Observation Errors

Assimilation requires relation of observations to $\mathbf{x}$

- observation model

$$\mathbf{y} = H(\mathbf{x}) + \epsilon$$

- H is observation (or *forward*) operator mapping state onto obs
- e.g., interpolation for in-situ obs of state variables or radiative-transfer integrals for remotely sensed radiances
- H may be nonlinear; errors may not be additive

Three sources of observation error

- measurement error: noise in instrument or uncertainty in location
- errors of representativeness: obs influenced by scales not represented in discrete basis of $\mathbf{x}$
- observation operator: relation of obs to $\mathbf{x}$ may be incorrectly specified
- ϵ is sum of all three effects

Multivariate Gaussians

Probability density function

- $p(\mathbf{x}) = c \exp\left((\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{P}^{-1} (\mathbf{x} - \bar{\mathbf{x}})\right)$
- completely specifed by mean $\bar{\mathbf{x}}$ and covaraince matrix $\mathbf{P}$
- $\mathbf{P}$ describes statistical relations between elements of $\mathbf{x}$

Standard assumptions

- Gaussian forecast errors: $\mathbf{x} \sim N(\mathbf{x}_b, \mathbf{B})$
- Gaussian obs errors, linear obs operator: $\mathbf{y} = \mathbf{H}\mathbf{x} + \epsilon$, with $\epsilon \sim N(\mathbf{0}, \mathbf{R})$

Bayes rule for Gaussians

- Products of Gaussian yield Gaussians, so posterior (or analysis) pdf is also Gaussian. Thus, need formulas for its mean and covariance
- analysis equations:

$$\mathbf{x}_a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{x}_b + \mathbf{K}\mathbf{y}_o \quad ; \quad \mathbf{A} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B},$$

- Kalman gain

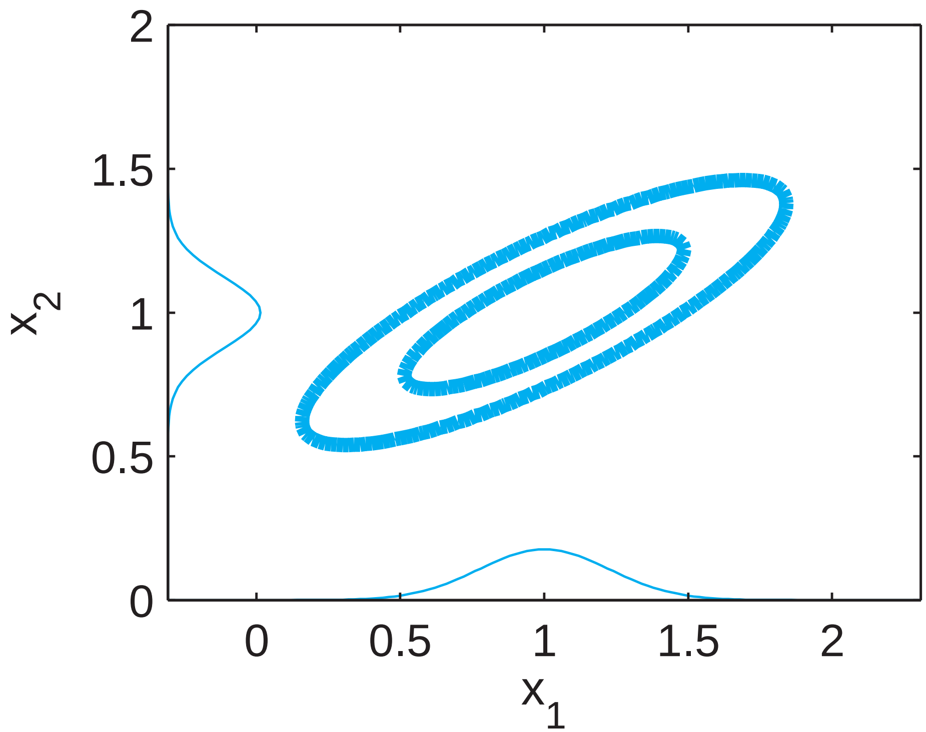
$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \quad [\text{c.f., } \alpha_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)\ ]$$

- equivalently, compute $\mathbf{x}_a$ as minimizer of

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b) + (\mathbf{y}_o - \mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}_o - \mathbf{H}\mathbf{x})^T$$

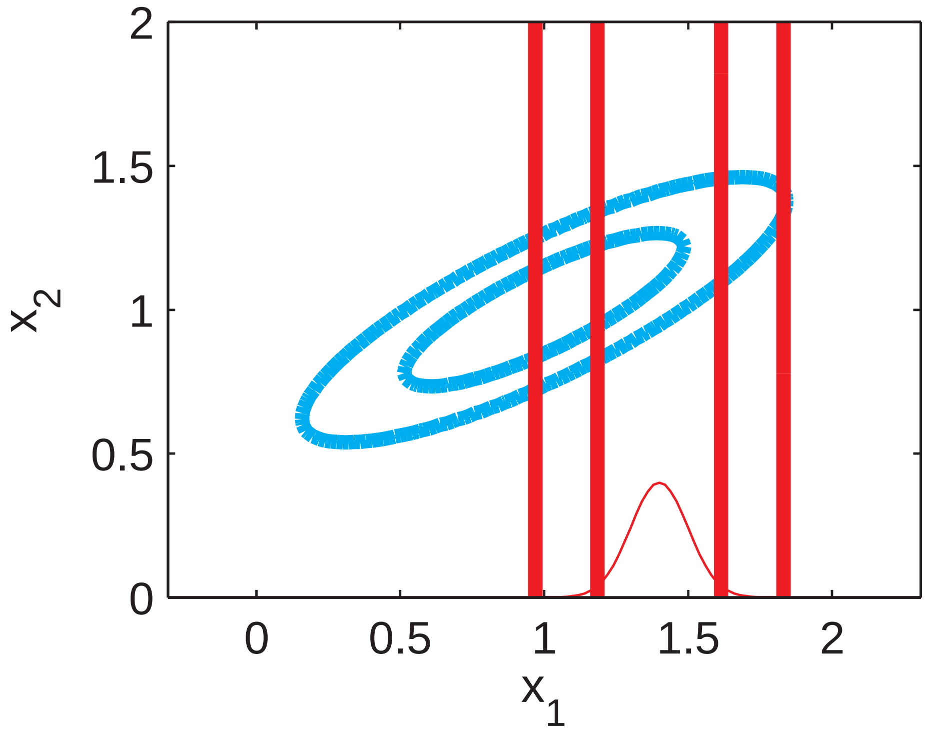
Importance of Covariances

- 2D: $\mathbf{x} = (x_1, x_2)$
- forecast/prior



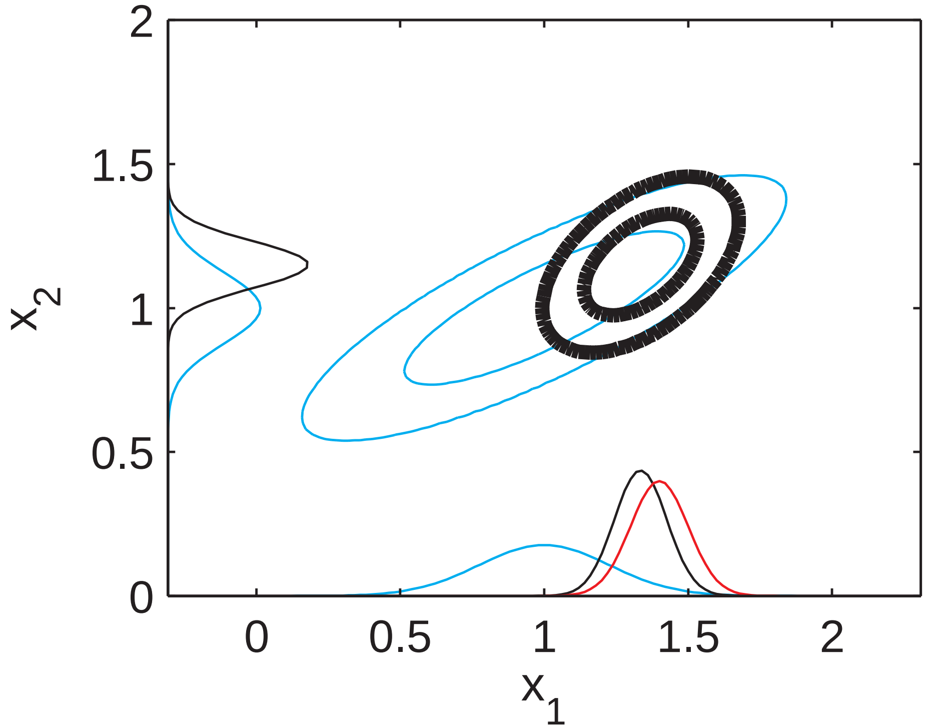
Importance of Covariances

- observation, $y = x_1^t +$ Gaussian noise $= 1.4$
- observation likelihood independent of x_2



Importance of Covariances

- analysis/posterior, $[x_1, x_2|y]$
- $\mathrm{Cov}(x_1, x_2)$ provides information on unobserved variable



Practical Considerations

Dimension of state vector may exceed 10^8

Implement Bayes rule in n dimensions

- pdfs are functions of n variables
- even discretizing each variable with 10 "points" requires 10^n total degrees of freedom

Implement Gaussian update

- densities specified by their mean and covariances
- requires n floating point numbers for mean, $n(n-1)/2$ for covariance
- moreover, background error covariances poorly known

Necessity for approximation and simplification

Additional Matters

Quality control

- identify observations with gross errors
- crucial in operational schemes

Computational issues

- tractable covariance models
- minimization

Propagation of information in time

- role of dynamical systems
- sources and evolution of forecast error

Nonlinearity and non-Gaussianity