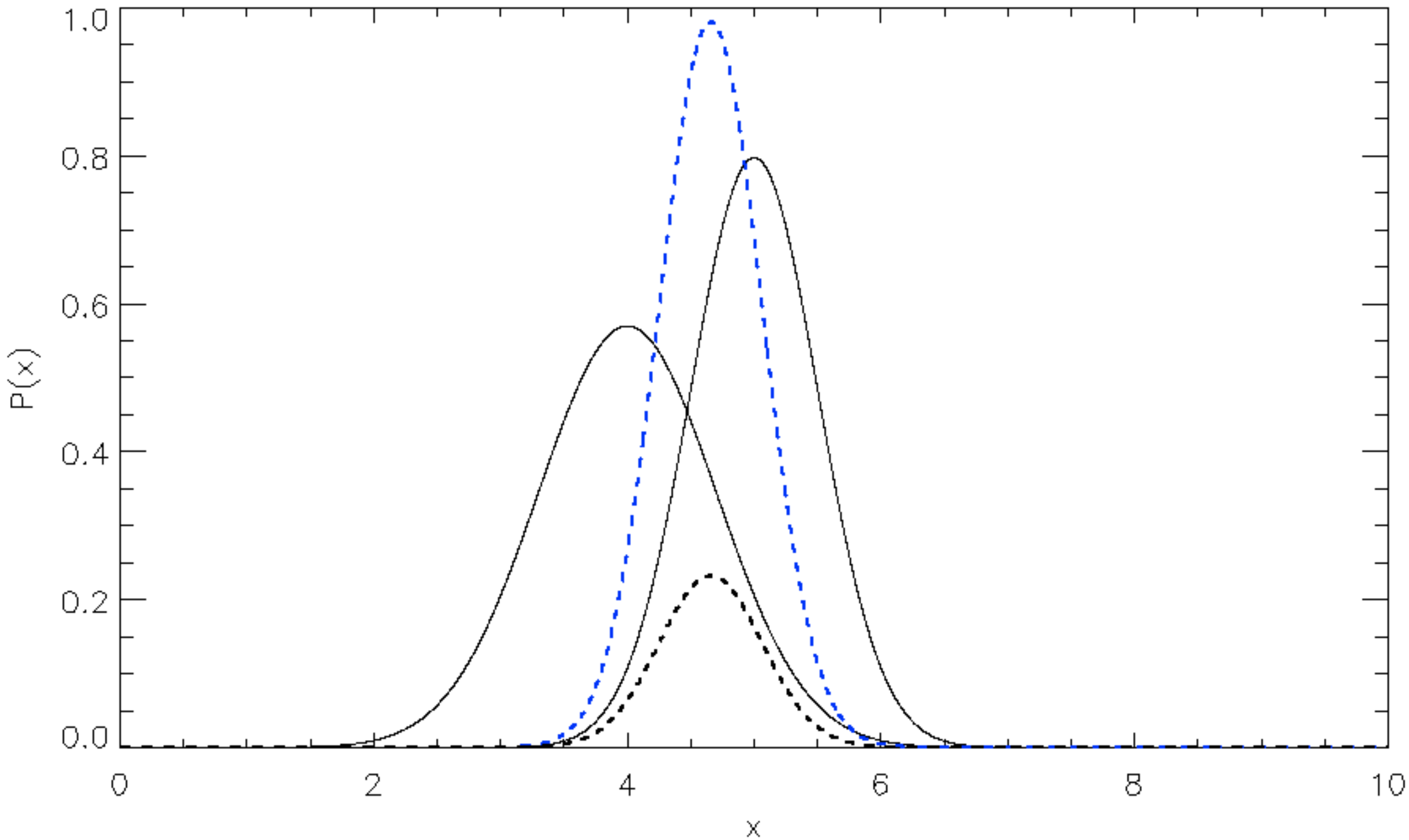
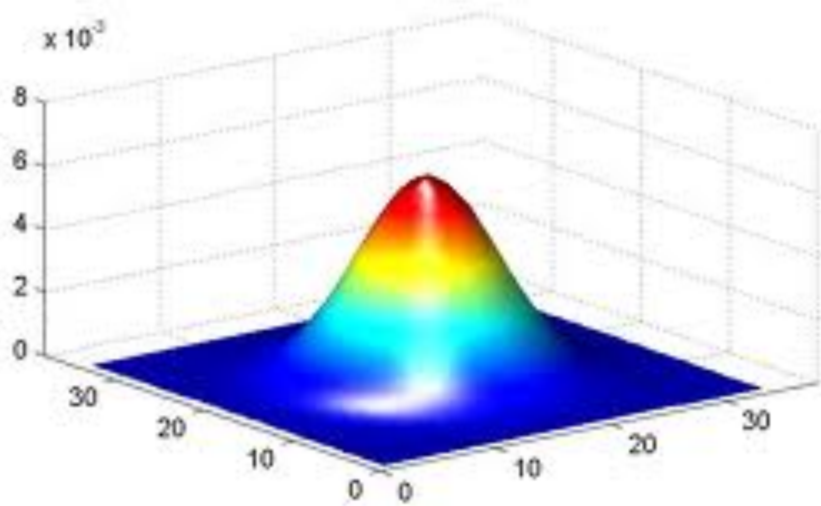
An Introduction to Optimal Estimation Theory

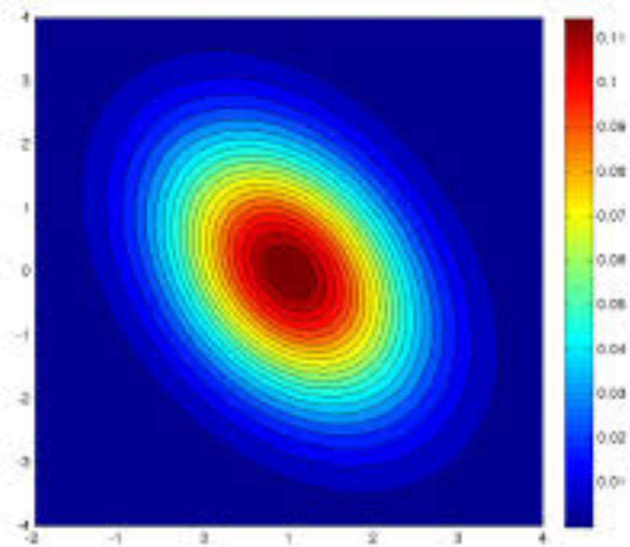
Chris O´Dell

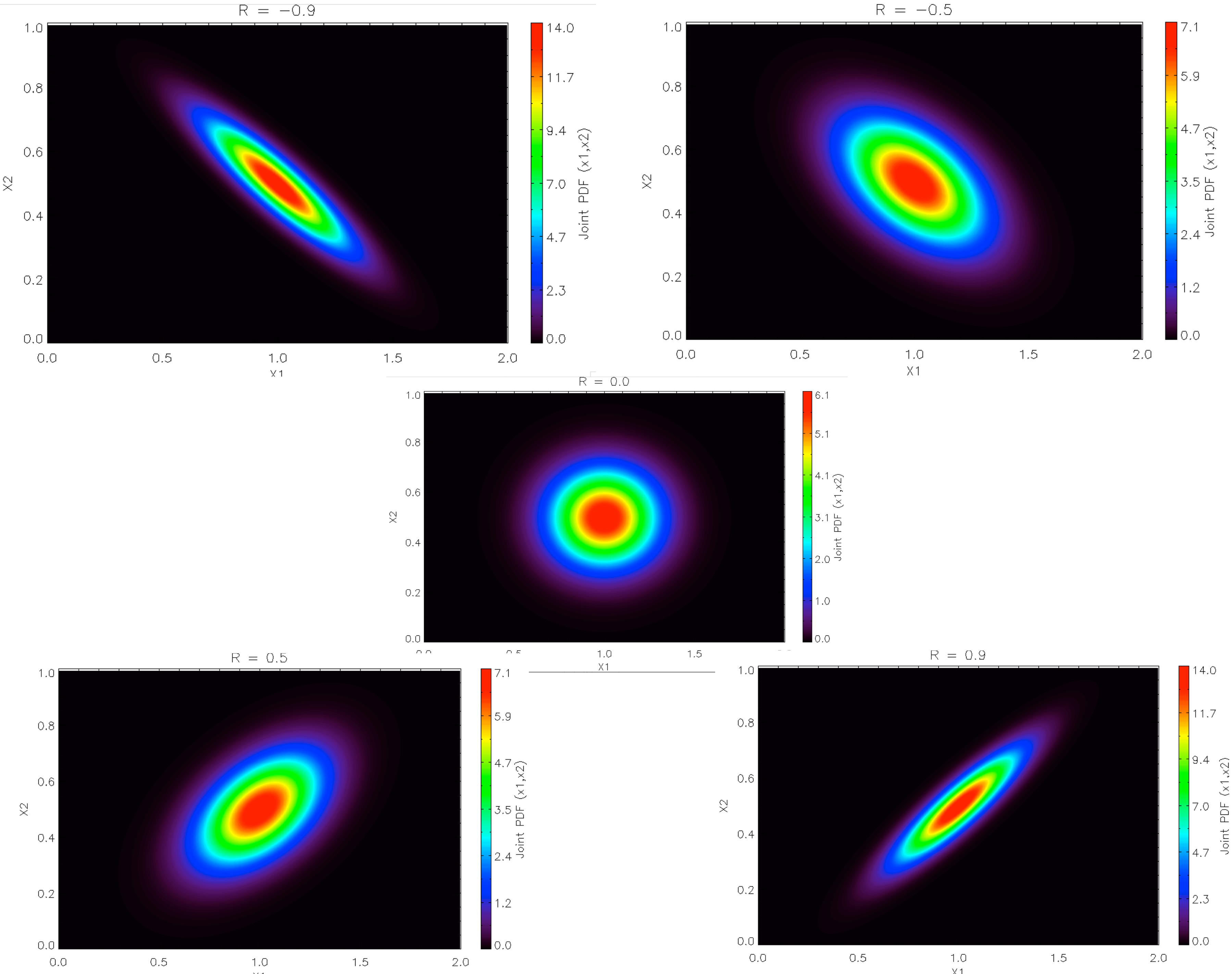
Updated April 2014


1.0
0.8
0.6
0.4
0.2
0.0
P(x)
0
2
4
6
8
10
x

2D Gaussian






R = −0.9
X2
X1
Joint PDF (x1,x2)
0.0 2.3 4.7 7.0 9.4 11.7 14.0
R = −0.5
0.0 1.2 2.4 3.5 4.7 5.9 7.1
R = 0.0
0.0 1.0 2.0 3.1 4.1 5.1 6.1
R = 0.5
0.0 1.2 2.4 3.5 4.7 5.9 7.1
R = 0.9
0.0 2.3 4.7 7.0 9.4 11.7 14.0
0.0 0.5 1.0 1.5 2.0
0.0 0.2 0.4 0.6 0.8 1.0

Bayes Theorem

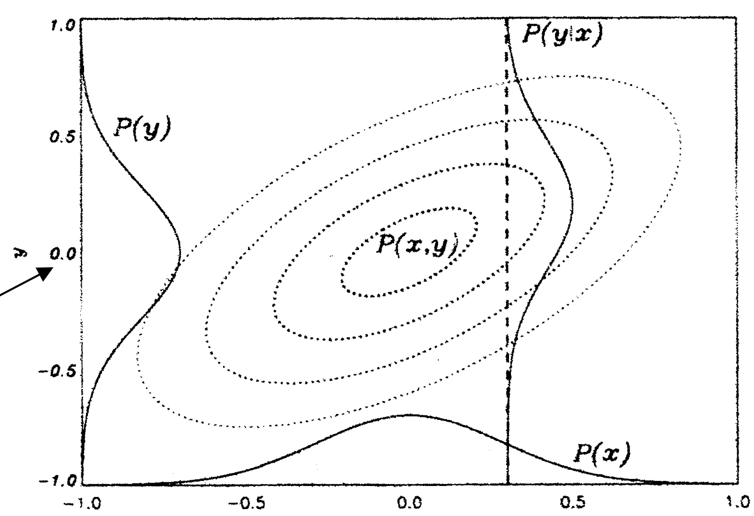
$P(\mathbf{x})d\mathbf{x}$ is the prior pdf of the state **x**. This means that the quantity P(**x**)d**x** is the probability (before the measurement) that **x** lies in the multidimensional volume (**x**,**x**+d**x**) expressing quantitatively our knowledge of **x** before the measurement is made.

$P(\mathbf{y})$ as the prior pdf of the measurement with a similar meaning. This is the pdf of the measurement *before it is made.* Typically taken to be flat or uniform.

$P(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}$ as the joint prior pdf of **x** and **y** meaning that is the probability that **x** lies in (**x**,**x**+d**x**) and **y** lies in (**y**,**y**+d**y**)

$P(\mathbf{y}|\mathbf{x})d\mathbf{y}$ as the conditional pdf of **y** given **x** meaning that is the probability that **y** lies in (**y**,**y**+d**y**) when **x** has a given value. This is a function derived by the forward model.

$P(\mathbf{x}|\mathbf{y})d\mathbf{x}$ as the conditional p
meaning that is the pro
in (**x**,**x**+d**x**) when **y** ha
This is the quantity of i
the inverse p



$$P(x) = \int_{-\infty}^{\infty} P(x,y)dy \text{ and } P(y) = \int_{-\infty}^{\infty} P(x,y)dx$$

Bayes Theorem

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}$$

This represents an updating to our prior knowledge $P(\mathbf{x})$ given the measurement $\mathbf{y}$

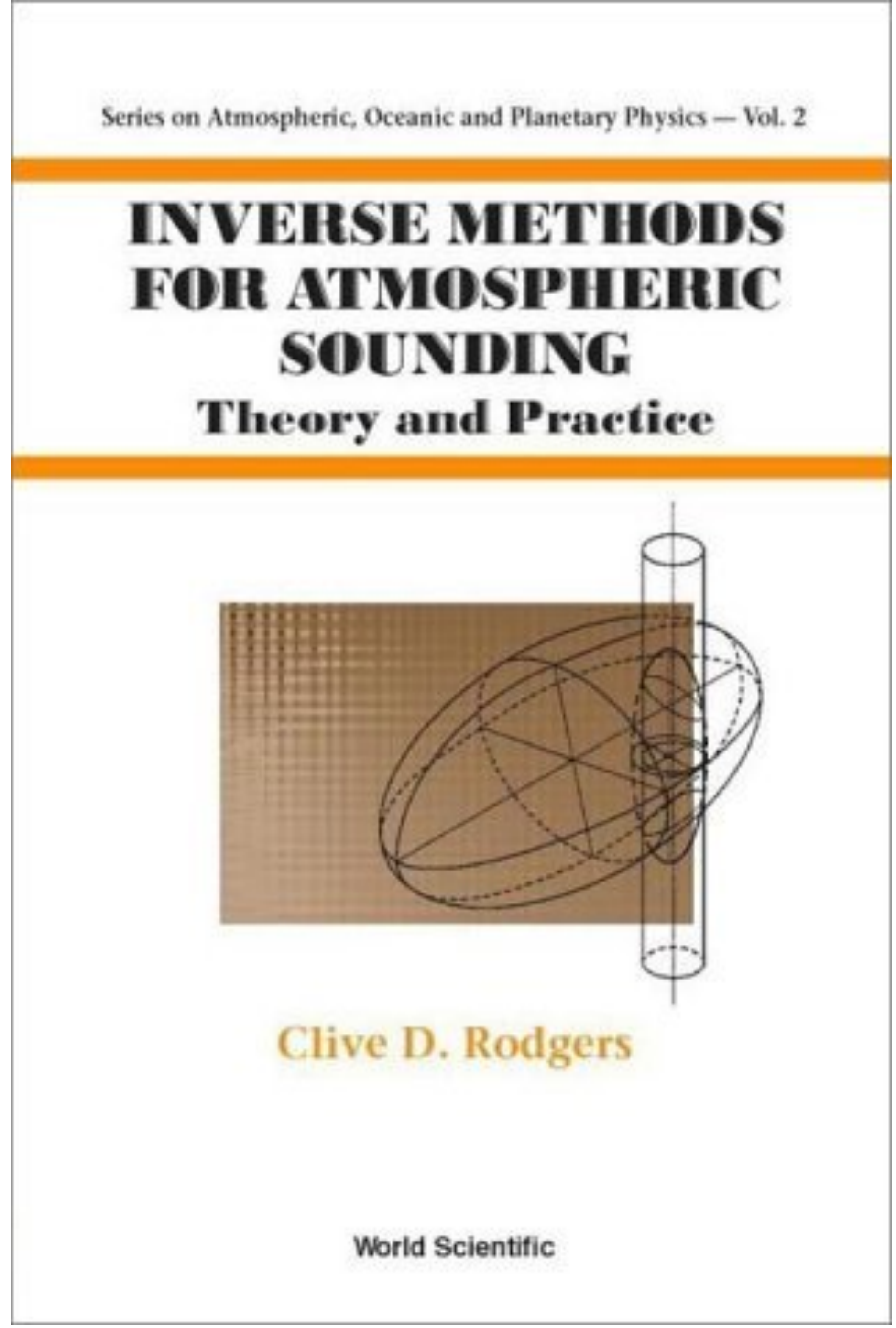
$P(\mathbf{y}|\mathbf{x})$ is the knowledge of $\mathbf{y}$ given $\mathbf{x}$ = pdf of forward model

The most likely value of $\mathbf{x}$ derived from this posterior pdf therefore represents our inverse solution.

Our knowledge contained in $P(\mathbf{y}|\mathbf{x})$ is explicitly expressed in terms of the forward model and the statistical description of both the error of this model and the error of the measurement.

The factor $P(\mathbf{y})$ will be ignored as it in practice is a normalizing factor.

The Bible of Optimal Estimation: ***Rodgers, 2000***



The Retrieval PROBLEM

DATA = FORWARD MODEL (State) + NOISE

OR

$$\vec{y} = \vec{F}(\vec{x}) + \vec{n}$$

DATA:

Reflectivities

Radiances

Radar Backscatter

Measured Rainfall

Et Cetera

Forward Model:

Cloud Microphysics

Precipitation

Surface Albedo

Radiative Transfer

Instrument Response

State Vector Variables:

Temperatures

Cloud Variables

Precip Quantities / Types

Measured Rainfall

Et Cetera

Noise:

Instrument Noise

Forward Model Error

Sub grid-scale processes

$$\vec{y} = \vec{F}(\vec{x}) + \vec{n}$$

We must *invert* this equation to get **x**, but how to do this:

• When there are different numbers of unknown variables as measured variables?

• In the presence of noise?

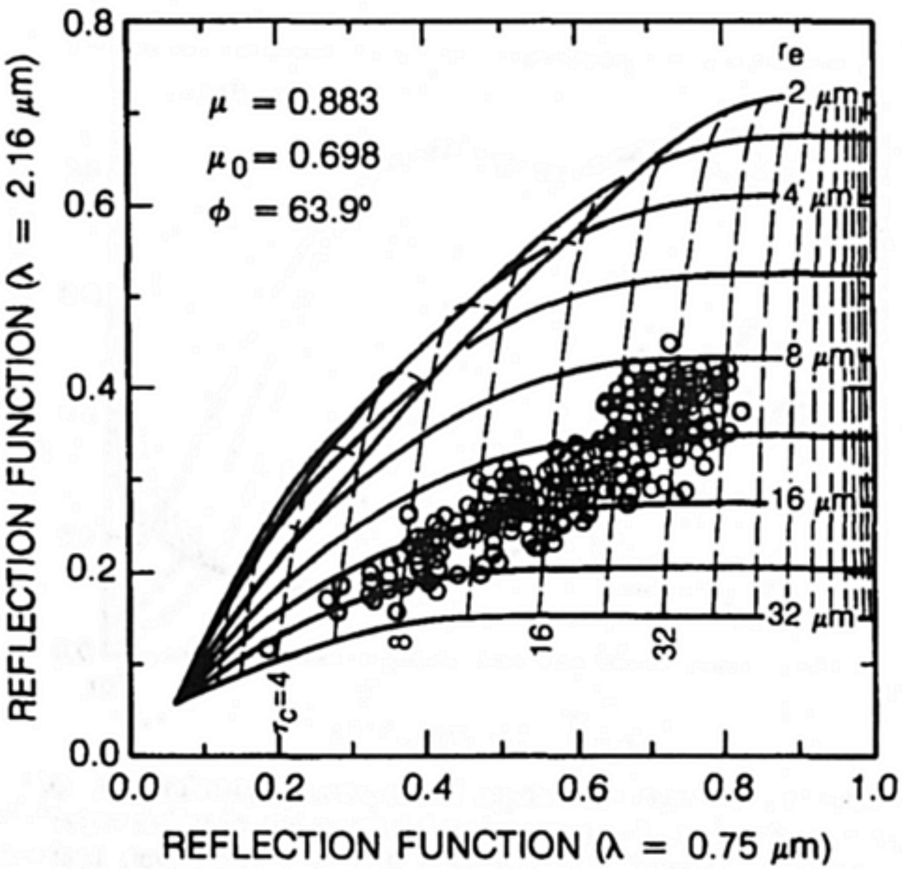
• When F(x) is nonlinear (often highly) ?

Retrieval Solution Techniques

- Direct Solution (requires invertible forward model and same number of measured quantities as unknowns)
- Multiple Linear Regression
- Semi-Empirical Fitting
- Neural Networks
- Optimal Estimation Theory

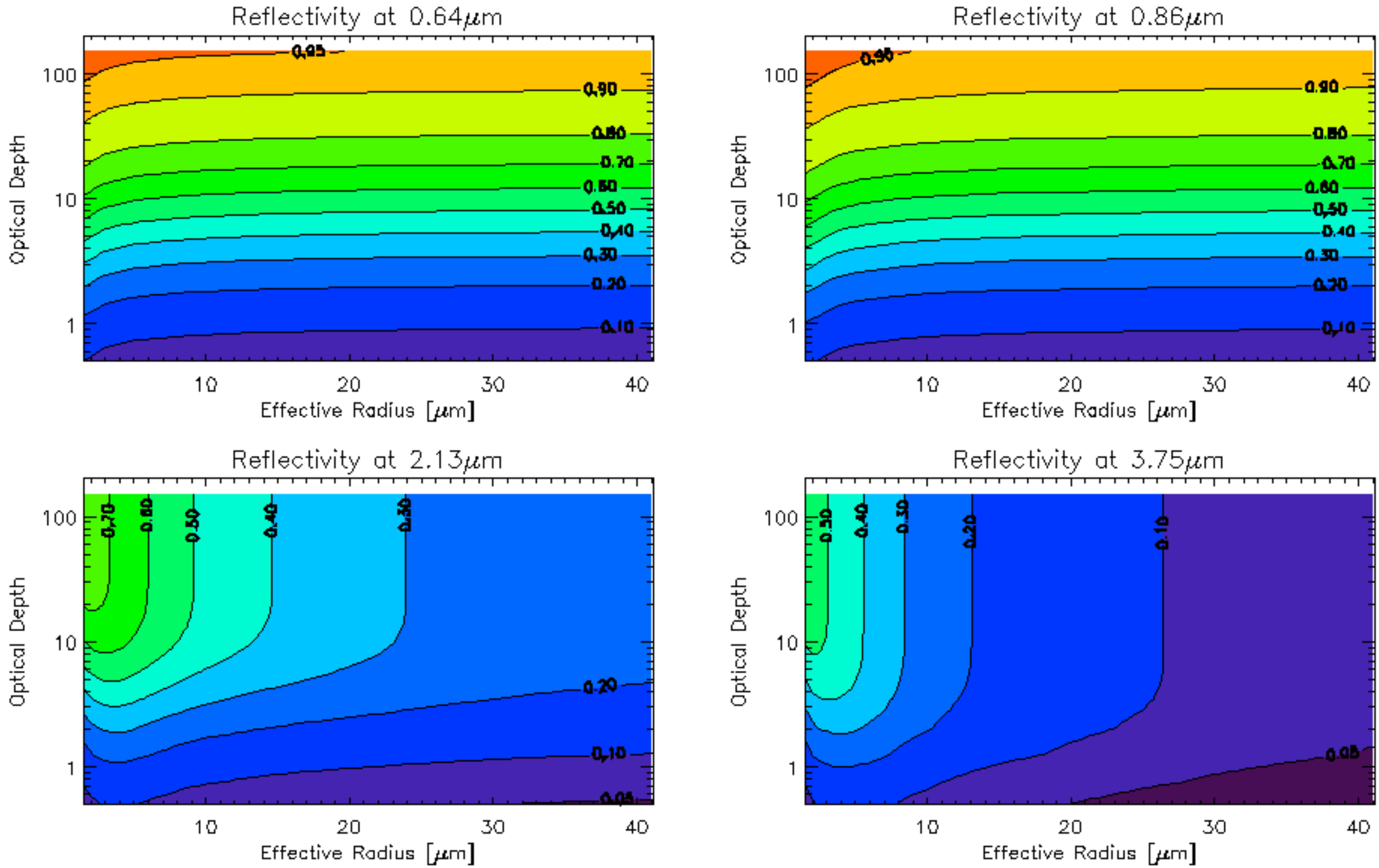
Technique	**Accuracy**	**Speed**	**Error Estimate?**	**Error Correlations?**	**Can use prior knowledge?**	**Information Content?**
Direct Solution	High if no errors, can be low	FAST	NO	NO	NO	NO
Linear Regression	Depends	FAST	Depends	NO	NO	NO
Neural Networks	Depends	MEDIUM	NO	NO	NO	NO
Optimal Estimation	Maximum	SLOW	YES	YES	YES	YES

A Simple Example: Cloud Optical Thickness and Effective Radius

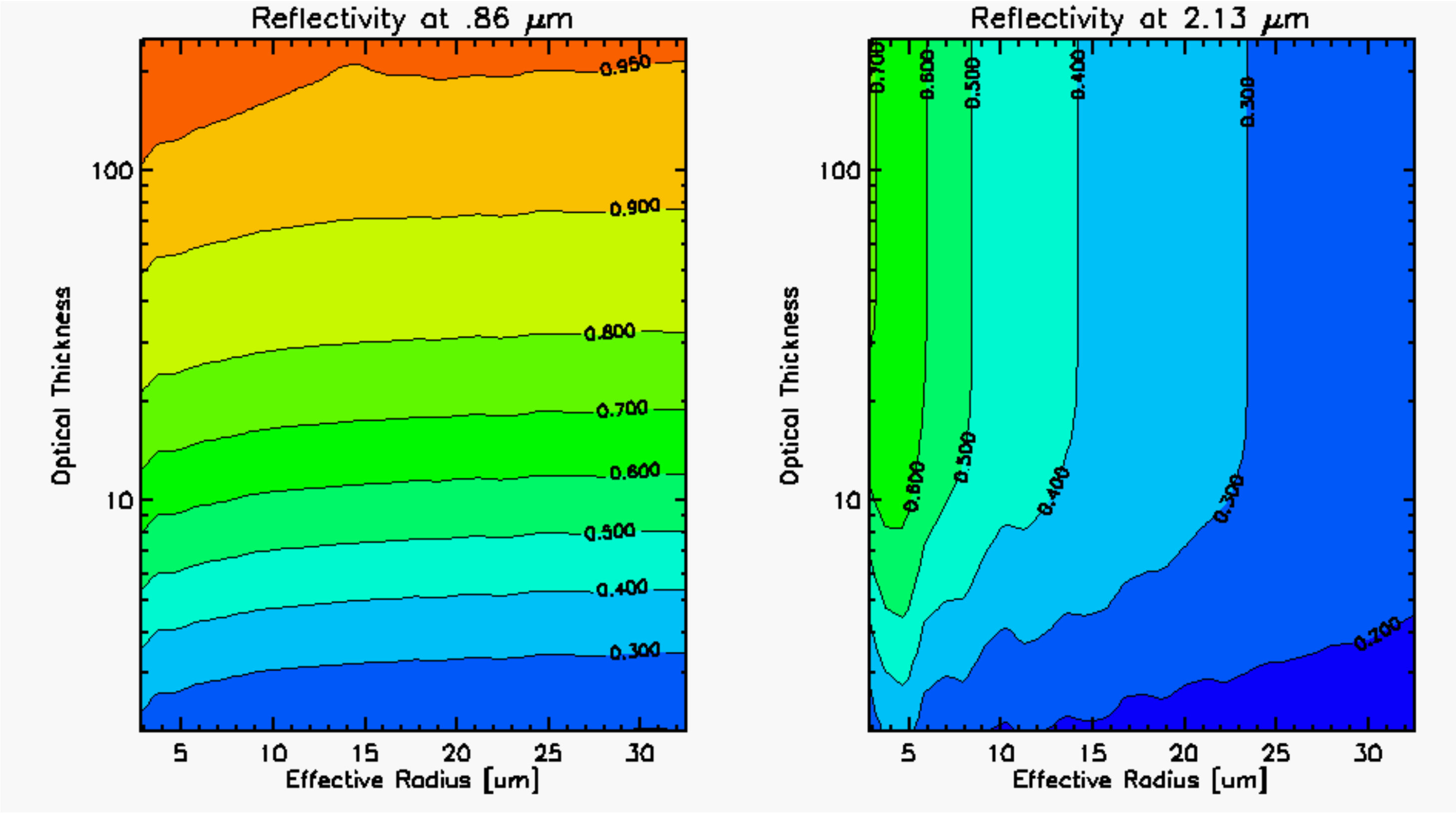


water cloud forward calculations from Nakajima and King 1990

- Cloud Optical Thickness and Effective Radius are often derived with a look-up table approach, although errors are not given.
- The inputs are typically 0.86 micron and 2.06 micron reflectances from the cloud top.



Reflectivity at 0.64μm
Reflectivity at 0.86μm
Reflectivity at 2.13μm
Reflectivity at 3.75μm
Optical Depth
Effective Radius [μm]



- $\mathbf{x} = \{\ r_{eff}, \tau\ \}$
- $\mathbf{y} = \{\ R_{0.86}\ , R_{2.13}\ ,\ \ldots\ \}$
- Forward model must map **x** to **y**. Mie Theory, simple cloud droplet size distribution, radiative transfer model.

Simple Solution:
The Look-up Table Approach

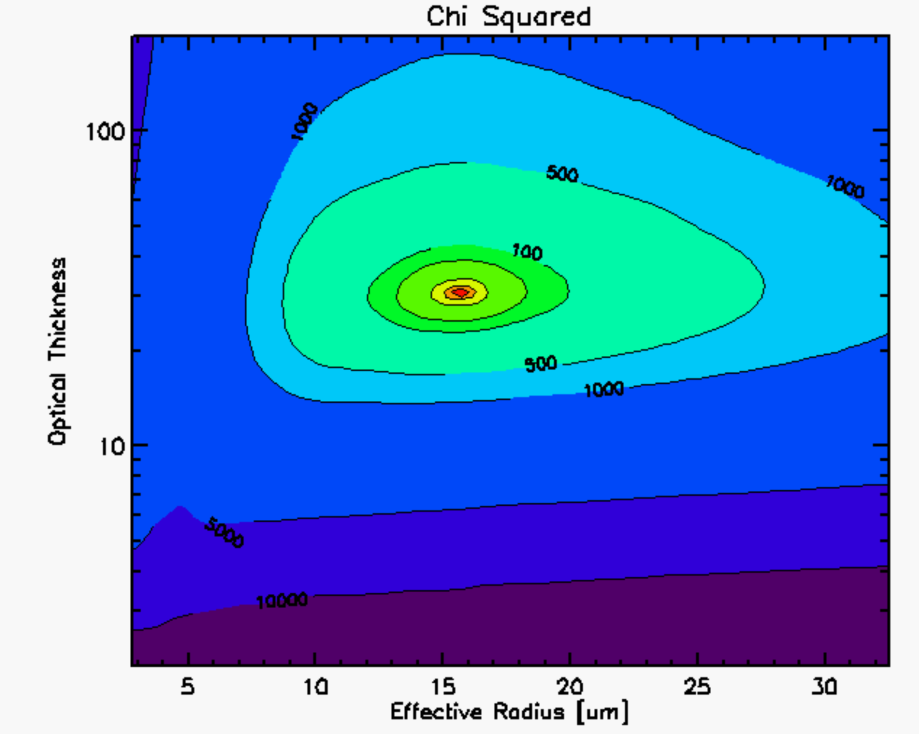
1. Assume we know $\vec{F}(\vec{x})$.
2. Regardless of noise, find $\hat{x}$ such that

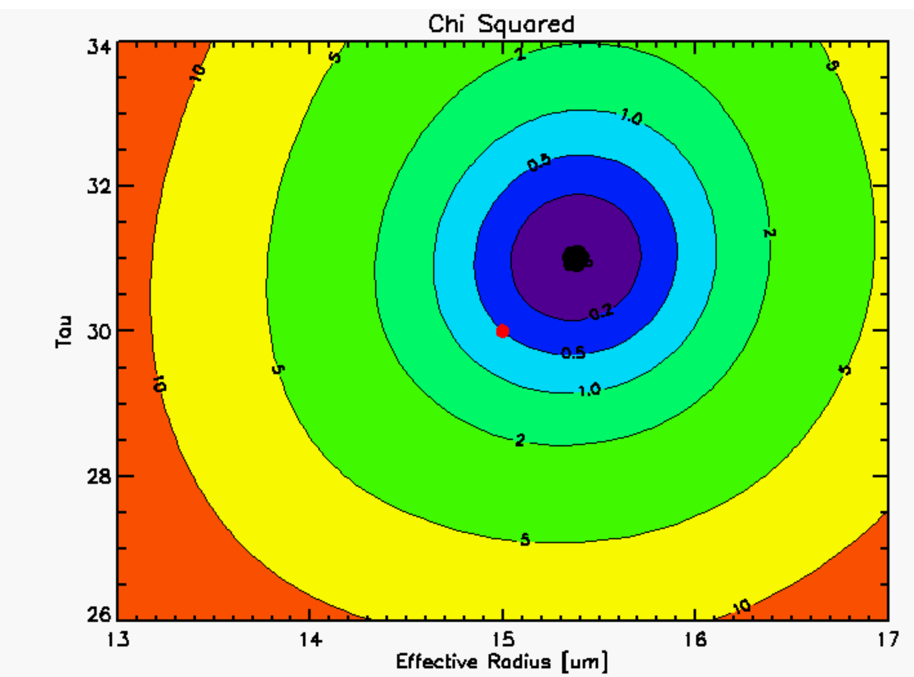
$$\vec{\varepsilon} = \vec{y} - \vec{F}(\vec{x})$$ is minimized.

3. But $\vec{\varepsilon}$ is a vector! Therefore, we minimize the χ^2 :

$$\chi^2 = \sum_i \frac{(y_i - F_i(\vec{x}))^2}{\sigma_i^2}$$

4. Can make a look-up table to speed up the search.





Example:

$\mathbf{x_{true}}$: { r_{eff} = 15 μm, τ = 30 }

Errors determined by how much change in each parameter (r_{eff}, τ) causes the χ^2 to change by one unit.

BUT

- What if the Look-Up Table is too big?
- What if the errors in **y** are correlated?
- How do we account for errors in the forward model?
- Shouldn’t the output errors be correlated as well?
- How do we incorporate *prior knowledge* about **x** ?

Correlated Errors

- Variables: y_1, y_2 , y_3 ...
- 1-sigma errors: σ_1 , σ_2 , σ_3 ...
- The correlation between y_1 and y_2

is c_{12} (between –1 and 1), etc.

- Then, the *Noise Covariance Matrix*

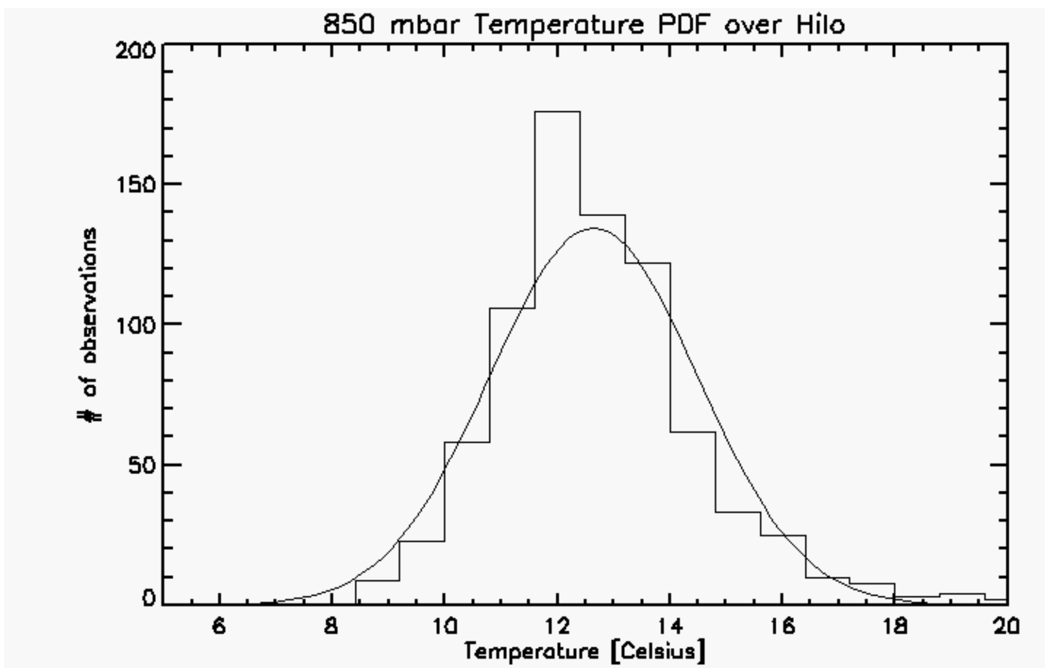
is given by:

$$S_y = \begin{bmatrix} \sigma_1^2 & c_{12}\sigma_1\sigma_2 & c_{13}\sigma_1\sigma_3 & \cdots \\ c_{12}\sigma_1\sigma_2 & \sigma_2^2 & c_{23}\sigma_2\sigma_3 & \cdots \\ c_{13}\sigma_1\sigma_3 & c_{23}\sigma_2\sigma_3 & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Example: Temperature Profile Climatology for December over Hilo, Hawaii

P = (1000, 850, 700, 500, 400, 300) mbar

<T> = (22.2, 12.6, 7.6, -7.7, -19.5, -34.1) Celsius



Correlation Matrix:

1.00	0.47	0.29	0.21	0.21	0.16
0.47	1.00	0.09	0.14	0.15	0.11
0.29	0.09	1.00	0.53	0.39	0.24
0.21	0.14	0.53	1.00	0.68	0.40
0.21	0.15	0.39	0.68	1.00	0.64
0.16	0.11	0.24	0.40	0.64	1.00

Covariance Matrix:

2.71	1.42	1.12	0.79	0.82	0.71
1.42	**3.42**	0.37	0.58	0.68	0.52
1.12	0.37	**5.31**	2.75	2.18	1.45
0.79	0.58	2.75	**5.07**	3.67	2.41
0.82	0.68	2.18	3.67	**5.81**	4.10
0.71	0.52	1.45	2.41	4.10	**7.09**

What to Minimize

- But what do we minimize ? Before, we found xhat such that this was a minimum:

$$\chi^2 = \sum_i \frac{(y_i - F_i(\vec{x}))^2}{\sigma_i^2}$$

- Now, we must minimize the *generalized* χ^2 :

$$\chi^2 = \left(\vec{y} - \vec{F}(\vec{x})\right)^T S_y^{-1} \left(\vec{y} - \vec{F}(\vec{x})\right)$$

- This is just a number (a scalar), and reduces to the first equation when all the correlations are *zero*.

Prior Knowledge

- Prior knowledge about $\mathbf{x}$ can be known from many different sources, like *other measurements* or *a weather or climate model prediction* or *climatology*.

- In order to specify prior knowledge of $\mathbf{x}$, called $\mathbf{x_a}$, we must also specify how well we know $\mathbf{x_a}$; we must specify the errors on $\mathbf{x_a}$.

- The errors on $\mathbf{x_a}$ are generally characterized by a Probability Distribution Function (PDF) with as many dimensions as $\mathbf{x}$.

- For simplicity, people often assume prior errors to be Gaussian; then we simply specify S_a, the error covariance matrix associated with $\mathbf{x_a}$.

The χ^2 with prior knowledge:

$$\chi^2 = \left(\vec{y} - \vec{F}(\vec{x}) \right)^T S_y^{-1} \left(\vec{y} - \vec{F}(\vec{x}) \right) + \left(\vec{x}_a - \vec{x} \right)^T S_a^{-1} \left(\vec{x}_a - \vec{x} \right)$$

Yes, that is a scary looking equation. But it is not so bad...

Minimization Techniques

Minimizing the χ^2 is *hard*. In general, you can use a look-up table (this still works, if you have tabulated values of **F**(**x**)), but if the lookup table approach is not feasible (i.e., it's too big), then you have to do *iteration*:

1. Pick a guess for **x**, called $\mathbf{x_0}$.
2. Calculate (or look up) $\mathbf{F}(\mathbf{x_0})$.
3. Calculate (or look up) the *Jacobian Matrix* about $\mathbf{x_0}$:

$$K_{ij}(\vec{x}) = \frac{\partial F_i(\vec{x})}{\partial x_j}$$

K is the matrix of sensitivities, or derivatives, of each output (y) variable with respect to each input (x) variable. It is not necessarily square.

4. Finally, iterate to get the next guess:

$$\vec{x}_{i+1} = \vec{x}_i + S_i \left(K_i^T \; S_y^{-1} \left(\vec{y} - \vec{F}(\vec{x}_i) \right) + S_a^{-1} \left(\vec{x}_a - \vec{x}_i \right) \right)$$

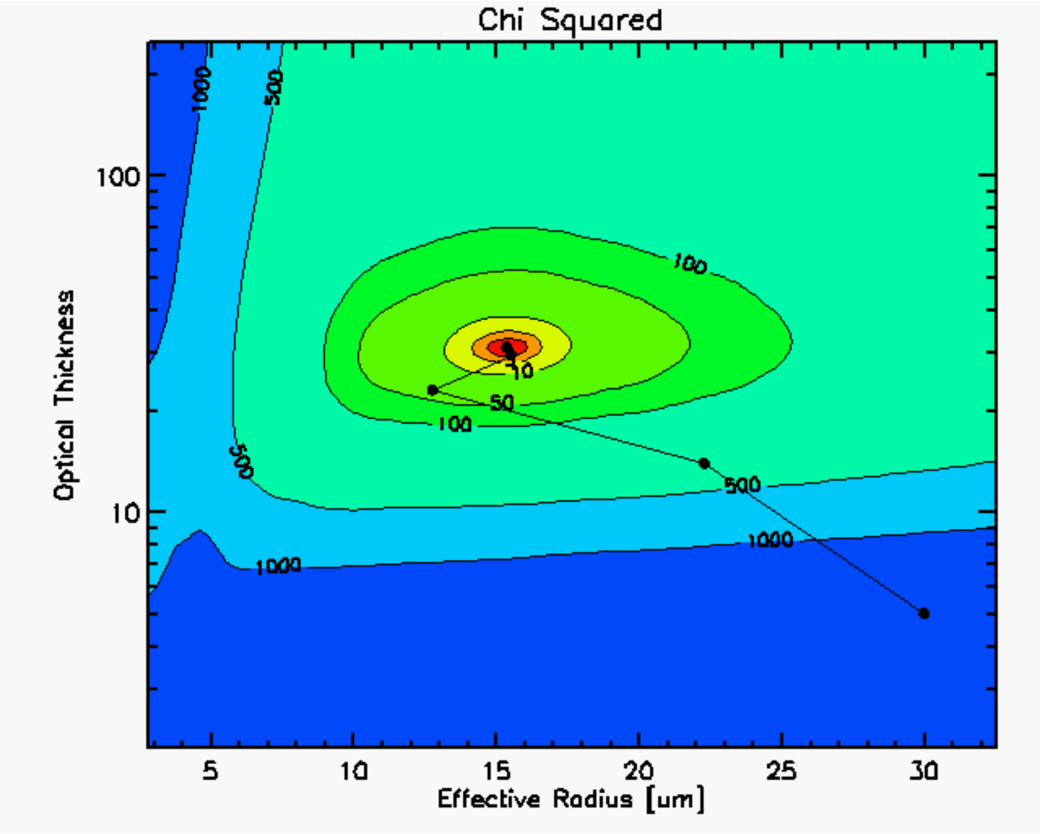
where:

$$S_i = \left(K_i^T \; S_y^{-1} K_i + S_a^{-1} \right)^{-1}$$

$$K_i = K(\vec{x}_i) = \frac{\partial \vec{F}(\vec{x}_i)}{\partial \vec{x}}$$

Iteration in practice:

- Not guaranteed to converge.
- Can be slow, depends on non-linearity of **F**(**x**).
- There are many tricks to make the iteration faster and more accurate.
- Often, only a few function iterations are necessary.



Error Correlations?

$\mathbf{x_{true}}$	r_{eff} = 15 μm, τ = 30	r_{eff} = 12 μm, τ = 8
$\{R_{0.86}, R_{2.13}\}_{true}$	0.796 , 0.388	0.516, 0.391
$\{R_{0.86}, R_{2.13}\}_{measured}$	0.808 , 0.401	0.529, 0.387
$\mathbf{x_{derived}}$	r_{eff} = 14.3 μm, τ = 32.3	r_{eff} = 11.8 μm, τ = 7.6
Formal 95% Errors	± 1.5 μm, ± 3.7	± 2.2 μm, ± 0.7
$\{r_{eff}, \tau\}$ **Correlation**	5%	55%

So, again, what is optimal estimation?

Optimal Estimation is a way to infer information about a system, based on observations. It is necessary to be able to simulate the observations, given complete knowledge of the system state.

Optimal Estimation can:

- Combine different observations of different types.

- Utilize prior knowledge of the system state (climatology, model forecast, etc).

- Errors are automatically provided, as are error correlations.

- Estimate the *information content* of additional measurements.

Applications of Optimal Estimation

- **Retrieval Theory** (standard, or using climatology, or using model forecast.   Combine radar & satellite. Combine multiple satellites. Etc.)

- **Data Assimilation** – optimally combine model forecast with measurements. Not just for weather! *Example: carbon source sink models, hydrology models.*

- **Channel Selection**: can determine information content of additional channels when retrieving certain variables. *Example: SIRICE mission is using this technique to select their IR channels to retrieve cirrus IWP.*