P.Weidelt

Inversion mit Vorinformation

Bezeichnungen:

Matrizen	$\underline{A}$, $\underline{B}$, $\underline{C}$,...
Vektoren	$\underline{a}$, $\underline{b}$, $\underline{c}$,...
Transposition	$\underline{a}^T$, $\underline{A}^T$

1. Einführung

Die Interpretation geophysikalischer Daten liefert selten eindeutige Ergebnisse. Selbst wenn die Eindeutigkeit theoretisch gesichert ist, führt die Unvollkommenheit jedes realen Datensatzes zu einer ganzen Schar von akzeptablen Lösungen. Die Stabilisierung der Lösung durch Zusatzinformation ist deshalb wünschenswert. In der Praxis wird häufig eine Stabilisierung dadurch erreicht, daß aufgrund bekannter oder vermuteter Eigenschaften des Modells das Inversionsproblem auf wenige Parameter reduziert wird (z.B. eindimensionales Modell mit einer vorgegebenen Anzahl homogener Schichten). Die Verwendung von Vorinformation ist notwendig und erstrebenswert. Wichtig ist nur, daß die Zuverlässigkeit der Vorinformation nicht überschätzt wird, da andernfalls das Interpretationsergebnis nur die falsch geschätzte Vorinformation widerspiegelt. In den beiden folgenden Beispielen ist eine sinnvolle Interpretation nur mit Zusatzinformation möglich:

a) Die grundsätzlich mehrdeutigen Potentialfelddaten (Gravimetrie, Magnetik) lassen sich vollständig durch eine äquivalente dünne Schicht unmittelbar unterhalb der Beobachtungsebene deuten. Für eine an die geophysikalische Situation angepaßte Interpretation sind deshalb Annahmen über die möglichen Störkörper erforderlich.

b) Das weniger triviale zweite Beispiel stammt aus der Magnetotellurik: Versucht man einen Satz gemessener frequenzabhängiger

elektromagnetischer Oberflächenimpedanzen $Z(\omega_j)$, $j = 1,\ldots, n$ durch eine geschichtete Leitfähigkeitsverteilung $\sigma(z)$ zu interpretieren, so findet man als bestpassendes Modell

$$\sigma(z) = \sum_k \tau_k \delta(z - z_k)$$

(Parker, 1980), d.h. $\sigma(z)$ ist entartet und besteht nur aus nichtleitenden Schichten und sehr gut leitenden dünnen Schichten mit der integrierten Leitfähigkeit τ_k in der Tiefe z_k. Für eine geophysikalisch sinnvolle Interpretation muß man daher unter Verzicht auf optimale Datenanpassung (zumindest implizit) Obergrenzen für Leitfähigkeiten und Untergrenzen für Schichtmächtigkeiten einführen.

Ein gezielter Einsatz von Vorinformation könnte bei den folgenden Problemen nützlich sein:

a) Interpretiert man etwa entlang einem Profil geophysikalische Daten durch eindimensionale Modelle, so zeigen derartige Interpretationen häufig schon zwischen benachbarten Stationen wenig plausibel erscheinende starke laterale Unterschiede. Man kann deshalb ein lateral korreliertes Tiefenprofil dadurch zu gewinnen versuchen, daß die Modellparameter und Streubreiten von Punkt A als Vorinformation zur Modellkonstruktion am benachbarten Punkt B benutzt werden.

b) Wenn in einem Gebiet verschiedene geowissenschaftliche Verfahren zum Einsatz kommen, die entweder auf denselben physikalischen Parameter oder auf verschiedene Parameter mit bekanntem oder vermutetem funktionalen Zusammenhang ansprechen, so besteht ein Fernziel der Interpretation darin, aus der Verknüpfung aller vorhandenen Informationen ein konsistentes Modell zu erstellen und Vorhersagen zu treffen (KTB). Man kann sich vorstellen, daß durch gezielte Verwendung von Vorinformation und neuen Daten der Wissensstand systematisch erweitert werden kann.

Die Verwendung von Vorinformation (oder a-priori-Information) zur Lösung des geophysikalischen Inversionsproblems hat in den letzten Jahren verstärktes Interesse gefunden. Wichtige Beiträge dazu stammen von Tarantola & Valette (1982a,b) und Jackson & Matsu'ura (1985). Eine gute Einführung gibt Menke (1984, p. 79-99 et 147-160). Die folgende Darstellung, die auf diese neuere Entwicklung aufmerksam machen soll, beruht im wesentlichen auf diesen Arbeiten.

2. Inversion mit Vorinformation: Definitionen und Ideen

Wir betrachten ein diskretes Problem, in dem das Modell durch den m-komponentigen Parametervektor $\underline{x}$,

$$\underline{x}^T = (x_1, \ldots, x_m) \tag{2.1}$$

beschrieben wird und zu dessen Bestimmung ein n-komponentiger Datenvektor $\underline{y}_o$,

$$\underline{y}_o^T = (y_{o1}, \ldots, y_{on}) \tag{2.2}$$

als fehlerbehafteter Zufallsvektor mit der bekannten Daten-Kovarianzmatrix $\underline{C}_y$ ("Streuung") zur Verfügung steht. Im einfachsten Fall ist

$$\underline{C}_y = \text{diag}\,(\sigma_{y1}^2, \ldots, \sigma_{yn}^2)\ . \tag{2.3}$$

Mit $\underline{x}$ verknüpft ist der n-komponentige Vektor der Datenfunktionale,

$$\underline{y} = \underline{f}(\underline{x}), \tag{2.4}$$

die die Regeln angeben, wie aus einem vorgegebenen Modell $\underline{x}$ der theoretische Datenvektor $\underline{y}$ erhalten werden kann.

Als Vorinformation für das Modell $\underline{x}$ wird nun angenommen, daß wir bereits eine Vorstellung über "vernünftige" Parameterwerte und ihre Streubreite besitzen. Diese Vorstellungen werden quantifiziert durch das a-priori-Modell $\underline{x}_o$ und die a-priori-Kovarianzmatrix $\underline{C}_x$. Im einfachsten Fall ist

$$\underline{C}_x = \text{diag}\,(\sigma_{x1}^2, \ldots, \sigma_{xm}^2). \qquad (2.5)$$

Die Größe der Komponenten von $\underline{C}_x$ beschreibt den Wissensstand. Bei schwacher Vorinformation sind alle Komponenten groß. - Unser Wissen über das Modell ist also gegeben durch $\underline{x}_o$, $\underline{C}_x$, $\underline{y}_o$, $\underline{C}_y$ und die Datenfunktionale (2.4).

Es sei $\tilde{\underline{x}}$ das wahre aber unbekannte Modell und $\tilde{\underline{y}} = \underline{f}(\tilde{\underline{x}})$ der zugehörige theoretische Datenvektor. Dann gilt

$$\underline{x}_o = \tilde{\underline{x}} + \underline{e}_x, \quad \underline{y}_o = \tilde{\underline{y}} + \underline{e}_y, \qquad (2.6)$$

wobei $\underline{e}_x$ und $\underline{e}_y$ die ebenfalls unbekannten Schätzfehler des a-priori-Modells $\underline{x}_o$ und unbekannten Datenfehler des Datenvektors $\underline{y}_o$ sind. Wir wollen für das Folgende annehmen, daß $\underline{e}_x$ und $\underline{e}_y$ durch m- bzw. n-dimensionale Normalverteilungen mit Mittelwert $\underline{0}$ und bekannter Kovarianzmatrix $\underline{C}_x$ und $\underline{C}_y$ dargestellt werden können,

$$\underline{e}_x = N(\underline{0}, \underline{C}_x), \quad \underline{e}_y = N(\underline{0}, \underline{C}_y).$$

Dann gehört zu $\underline{e}_x = \underline{x}_o - \tilde{\underline{x}}$ die m-dimensionale <u>Wahrscheinlichkeitsdichtefunktion</u> (Wdf)

$$p(\underline{x}) = \frac{1}{\left[(2\pi)^m \det \underline{C}_x\right]^{1/2}} \exp\left\{-\tfrac{1}{2}(\underline{x}-\underline{x}_o)^T \underline{C}_x^{-1} (\underline{x}-\underline{x}_o)\right\}. \qquad (2.7a)$$

Der im übrigen unwichtige Vorfaktor sorgt lediglich dafür, daß

$$\int p(\underline{x})\, d\underline{x} = 1.$$

Dabei ist $d\underline{x}$ das m-dimensionale Volumenelement, die Integration erstreckt sich über den ganzen R^m.

Im wichtigen Spezialfall (2.5) ist $p(\underline{x})$ einfach das Produkt von m eindimensionalen Gaußverteilungen. Für die Verteilung von $\underline{y}$ gilt entsprechend

$$p(\underline{y}) = \frac{1}{\left[(2\pi)^n \det \underline{C}_y\right]^{1/2}} \exp\left\{-\tfrac{1}{2}(\underline{y}-\underline{y}_o)^T \underline{C}_y^{-1} (\underline{y}-\underline{y}_o)\right\}. \qquad (2.7b)$$

$p(\underline{x})$ ist die <u>a-priori-Wdf</u> von $\underline{x}$, die von den Beobachtungen $\underline{y}_o$ vollkommen unabhängig ist. Nach Berücksichtigung von $\underline{y}_o$ erhält man aus $p(\underline{x})$ die <u>a-posteriori-Wdf</u> $p(\underline{x}|\underline{y}_o)$, die nun bestimmt werden soll.

Zwei leicht unterschiedliche Argumentationen mit identischen Ergebnissen bieten sich an. Die erste benutzt den Satz von Bayes: Es sei

a) $p(\underline{x})$ die a-priori-Wdf von $\underline{x}$, d.h. (2.7a),

b) $p(\underline{y}_o|\underline{x})$ die bedingte Wdf von $\underline{y}_o$ wenn $\underline{x}$ gegeben ist, d.h. nach (2.7b) mit $\underline{y} = \underline{f}(\underline{x})$

$$p(\underline{y}_o|\underline{x}) = \frac{1}{[(2\pi)^n \det \underline{C}_y]^{1/2}} \exp\{-\tfrac{1}{2}(\underline{y}_o-\underline{f}(\underline{x}))^T \underline{C}_y^{-1} (\underline{y}_o-\underline{f}(\underline{x}))\}, \quad (2.8)$$

c) $p(y_o)$ die unbedingte Wdf von $\underline{y}_o$, d.h.

$$p(\underline{y}_o) = \int p(\underline{y}_o|\underline{x})\, p(\underline{x})\, d\underline{x}, \quad (2.9)$$

d) $p(\underline{x}|\underline{y}_o)$ die a-posteriori-Wdf von $\underline{x}$ wenn $\underline{y}_o$ gegeben ist.

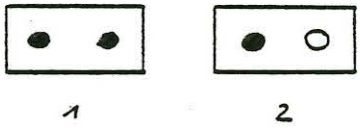
Dann lautet der Satz von Bayes:

$$\boxed{p(\underline{x}|\underline{y}_o) = p(\underline{y}_o|\underline{x})\, p(\underline{x})/p(\underline{y}_o)} \quad (2.10)$$

da die (m+n)-dimensionale Wdf $p(x, y_o)$ für das Auftreten von $\underline{x}$ und $\underline{y}_o$ gegeben ist durch

$$p(\underline{x},\underline{y}_o) = p(\underline{x}|\underline{y}_o)\, p(\underline{y}_o) = p(\underline{y}_o|\underline{x})\, p(\underline{x}).$$

Das obige Theorem des englischen Pastors und Mathematikers Thomas Bayes (1702-1761) wurde posthum in einer Arbeit mit dem Titel "An essay towards solving a problem in the doctrine of chances" 1763 veröffentlicht. Seine Bedeutung ist erst in neuerer Zeit erkannt worden. - Hier ein Trivialbeispiel für seine Anwendung im Falle diskreter Variabler: Gegeben seien zwei Urnen , die erste enthalte zwei schwarze Kugeln, die zweite eine schwarze und eine weiße. Aus einer beliebig herausgegriffenen Urne werde eine schwarze Kugel gezogen. Mit welcher Wahrscheinlichkeit stammt sie aus Urne 1 ? Diese Wahrscheinlichkeit ist offenbar 2/3. Mit dem Satz von Bayes gewinnt man sie so: x_i, i = 1,2 bedeute das Ereignis, daß die Urne i gewählt wird und y_o^1 stehe für eine schwarze Kugel. Dann gilt



$p(x_1) = p(x_2) = 1/2$, $p(y_o|x_1) = 1$, $p(y_o|x_2) = 1/2$,

$p(y_o) = p(y_o|x_1)\, p(x_1) + p(y_o|x_2)\, p(x_2) = 3/4$,

$p(x_1|y_o) = p(y_o|x_1)\, p(x_1)/p(y_o) = 2/3$. ■

Damit liefert (2.10) unter Fortlassung von Faktoren, die von $\underline{x}$ unabhängig sind, als a-posteriori-Wdf von $\underline{x}$

$$\boxed{p(\underline{x}|\underline{y}_o) \sim \exp\left\{-\frac{1}{2}(\underline{y}_o-\underline{f}(\underline{x}))^T\underline{C}_y^{-1}(\underline{y}_o-\underline{f}(\underline{x})) - \frac{1}{2}(\underline{x}-\underline{x}_o)^T\underline{C}_x^{-1}(\underline{x}-\underline{x}_o)\right\}} \qquad (2.11)$$

Der Satz von Bayes in der Form (2.11) (Annahme von normalverteilten Fehlern !) bildet die Grundlage für die Grundlage für die Verarbeitung von Vorinformation bei der Inversion.

Aus der Verteilung (2.11) wählt man als Schätzwert von $\tilde{\underline{x}}$ den Wert $\hat{\underline{x}}$, für den $p(\underline{x}|\underline{y}_o)$ sein Maximum annimmt (Prinzip der Maximum Likelihood). Die Notwendige Bedingung dafür gewinnt man aus (2.11) durch Differentiation nach $\underline{x}$:

$$\boxed{\underline{A}^T(\hat{\underline{x}})\underline{C}_y^{-1}\left\{\underline{f}(\hat{\underline{x}})-\underline{y}_o\right\} + \underline{C}_x^{-1}(\hat{\underline{x}}-\underline{x}_o) = 0} \qquad (2.12)$$

Dabei ist $\underline{A}(\hat{\underline{x}})$ die (n x m)-Jacobi-Matrix mit den Elementen

$$A_{ik}(\hat{\underline{x}}) = \left.\frac{\partial f_i}{\partial x_k}\right|_{\underline{x}=\hat{\underline{x}}} \qquad (2.13)$$

Bei der Herleitung von (2.12) wurde benutzt, daß mit einem m-Vektor $\underline{a}$ und einem n-Vektor $\underline{b}$ gilt

$$\nabla(\underline{a}^T\underline{x}) = \nabla(\underline{x}^T\underline{a}) = \underline{a}, \quad \nabla(\underline{b}^T\underline{f}) = \nabla(\underline{f}^T\underline{b}) = \underline{A}^T\underline{b}. \qquad (2.14)$$

Für nichtlineare Funktionale $\underline{f}(\underline{x})$ kann (2.12) i.a. nur iterativ gelöst werden (cf. Abschnitt 4).

Es folgt ein sehr einfaches, aber illustratives Beispiel zur Anwendung von (2.11) und (2.12):

Es sei m = 1, d.h. es soll eine skalare Größe x bestimmt werden, für die wir als Vorinformation die Schätzung x_o und deren Varianz $C_x = \sigma_x^2$ kennen mögen. Zur genaueren Bestimmung von x werden n Messungen mit den Ergebnissen y_{oi}, i = 1, ..., n durchgeführt. Die Datenfehler seien unkorreliert und es gelte $\underline{C}_y = \sigma_y^2 \underline{I}_n$. Was läßt sich nach Durchführung der Messungen über x und seine Streuung sagen ?

Es liegt offenbar das einfachste lineare Problem mit der "Theorie"

$$y_i = x, \quad i = 1, \ldots, n$$

vor. Die in (2.10) auftretenden Verteilungen sind

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{1}{2}\left(\frac{x-x_o}{\sigma_x}\right)^2\right\}, \tag{2.15a}$$

$$p(\underline{y}_o|x) = \frac{1}{(\sqrt{2\pi}\sigma_y)^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_{oi}-x}{\sigma_y}\right)^2\right\}, \tag{2.15b}$$

$$p(x|\underline{y}_o) = \frac{1}{\sqrt{2\pi}\,\sigma_{\hat{x}}} \exp\left\{-\frac{1}{2}\left(\frac{x-\hat{x}}{\sigma_{\hat{x}}}\right)^2\right\} \tag{2.15c}$$

mit

$$\frac{1}{\sigma_{\hat{x}}^2} = \frac{1}{\sigma_x^2} + \frac{n}{\sigma_y^2}, \quad \frac{\hat{x}}{\sigma_{\hat{x}}^2} = \frac{x_o}{\sigma_x^2} + \frac{\sum_{i=1}^{n} y_{oi}}{\sigma_y^2}. \tag{2.16}$$

Man gewinnt (2.15c) mit (2.10) aus (2.15a,b) durch "quadratische Ergänzung" der Summe der Exponenten und Bestimmung des (unwichtigen) Vorfaktors durch

$$\int p(x|\underline{y}_o)\, dx = 1.$$

$p(x|\underline{y}_o)$ ist also wieder eine Gaußverteilung.

Gl. (2.16) zeigt sehr klar, wie sich in $\hat{x}$ und $\sigma_{\hat{x}}$ Vorinformation (x_o, σ_x) und Messung ($\underline{y}_o$, σ_y) überlagern. Bei starker Vorinformation ($\sigma_x \to 0$) ist $\hat{x} = x_o$ und $\sigma_{\hat{x}} = \sigma_x$; schwache Vorinformtion ($\sigma_x \to \infty$) liefert das zu erwartende Ergebnis ("Fehlerfortpflanzungsgesetz")

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} y_{oi}, \quad \sigma_{\hat{x}} = \sigma_y/\sqrt{n}.$$

In diesem Beispiel wird der Wissensstand sprunghaft vom Anfangszustand (x_o, σ_x) in den nach n Messungen erreichten Endzustand ($\hat{x}$, $\sigma_{\hat{x}}$) überführt. Dieser Vorgang hätte jedoch auch als ein schrittweiser Lernprozess erfolgen können, indem für k = 1, ..., n-1 die a-posteriori-Wdf nach der k-ten Messung als a-priori-Wdf (x_{ok}, σ_{xk}) vor der (k+1)-ten Messung verwendet wird. Denn ausgehend von $x_{oo} := x_o$ und $\sigma_{xo} := \sigma_x$ ist nach (2.16) für k = 1, ..., n

$$\frac{1}{\sigma_{xk}^2} = \frac{1}{\sigma_{x,k-1}^2} + \frac{1}{\sigma_y^2} = \frac{1}{\sigma_x^2} + \frac{k}{\sigma_y^2}, \tag{2.17a}$$

$$\frac{x_{ok}}{\sigma_{xk}^2} = \frac{x_{o,k-1}}{\sigma_{x,k-1}^2} + \frac{y_{ok}}{\sigma_y^2} = \frac{x_o}{\sigma_x^2} + \frac{\sum_{i=1}^{k} y_{oi}}{\sigma_y^2}. \tag{2.17b}$$

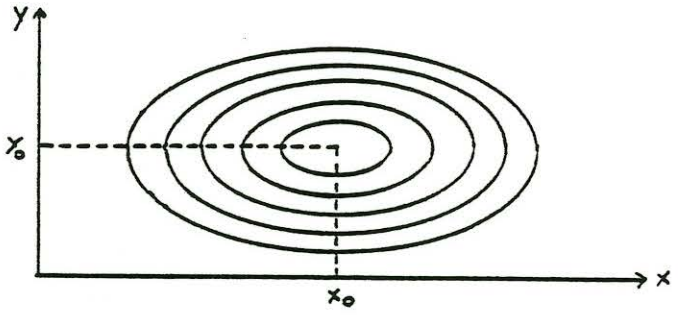
Es soll nun noch kurz auf die zweite Argumentation zur Herleitung von (2.12) eingegangen werden. Diese Argumentation betont besonders die Gleichartigkeit von Vorinformation und Daten. - Solange der Modellzusammenhang $\underline{y} = \underline{f}(\underline{x})$ unberücksichtigt bleibt, sind die Wdfs. (2.7a,b) voneinander unabhängig und können durch

$$p(\underline{x},\underline{y}) = p(\underline{x})\ p(\underline{y}) \tag{2.18}$$

zu einer (m+n)-dimensionalen Wdf zusammengefaßt werden. Die Größe $p(\underline{x},\underline{y})\ d\underline{x}\ d\underline{y}$ ist die Wahrscheinlichkeit dafür, daß das wahre Modell $\underline{x}$ zwischen $\underline{x}$ und $\underline{x}+ d\underline{x}$ und die wahren Daten $\underline{y}$ zwischen $\underline{y}$ und $\underline{y} + d\underline{y}$ liegen. Da aber $\underline{x}$ und $\underline{y}$ nicht unabhängig sind, sondern durch $\underline{y} = \underline{f}(\underline{x})$ miteinander verknüpft sind, interessiert nur der Wertevorrat von $p(\underline{x},\underline{y})$ für $\underline{y} = \underline{f}(\underline{x})$:

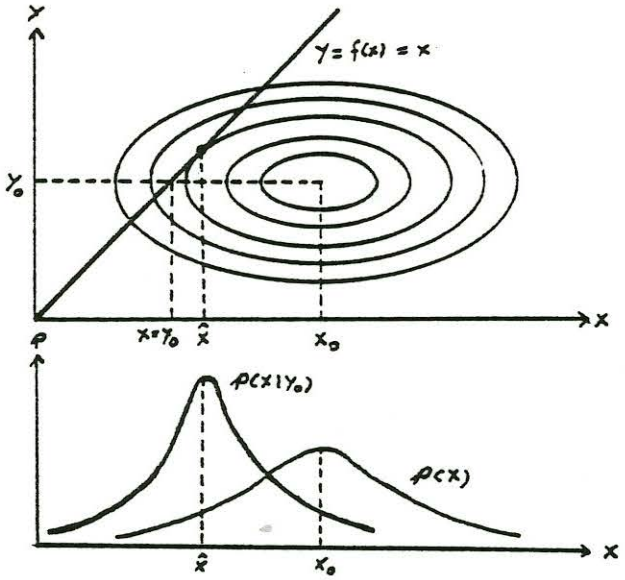
$$p(\underline{x},\underline{y}=\underline{f}(\underline{x})) = p(\underline{x})\ p(\underline{y}=\underline{f}(\underline{x})) = p(\underline{x})\ p(\underline{y}_o|\underline{x}) = p(\underline{x}|\underline{y}_o)\ p(\underline{y}_o) \tag{2.19}$$

Benutzt wurde (2.18), (2.7b) und (2.10). Die Wdf (2.18) für $\underline{y} = \underline{f}(\underline{x})$ ist damit proportional zur gesuchten a-posteriori Wdf von $\underline{x}$. Als Schätzwert $\hat{\underline{x}}$ von $\tilde{\underline{x}}$ wählt man den Punkt, für den $p(\underline{x},\underline{y})$ für $\underline{y} = \underline{f}(\underline{x})$ ein Maximum annimmt.- Für die hier vorausgesetzten Normalverteilungen sind die Flächen $p(\underline{x},\underline{y}) = \text{const}$ (m+n)-dimensionale Hyperellipsoide mit dem Mittelpunkt in $(\underline{x}_o,\underline{y}_o)$. Im Fall m = n = 1 können sie bei relativ zu den Daten schwacher Vorinformation etwa so aussehen:

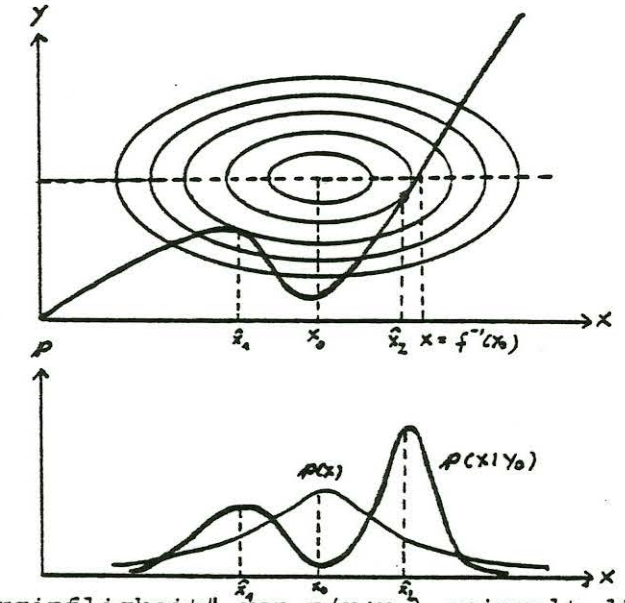


Diese Figur berücksichtigt noch nicht den theoretischen Zusammenhang zwischen Modell und Daten, $y = f(x)$. Wählt man wieder den einfachen Fall $y = x$, so ergibt sich die folgende Abbildung. In diesem linearen Fall ist $\hat{x}$ als Berührungspunkt der Geraden $y = x$ mit einer Ellipse eindeutig festgelegt. Wegen der schwachen Vorinformation liegt $\hat{x}$ nahe dem Punkt $x = y_o$, der sich ohne Vorwissen er-

geben hätte. Die Skizze zeigt im unteren Teil auch den prinzipiellen Verlauf von a-priori Wdf $p(x)$ und a-posteriori Wdf $p(x|y_o)$, wobei letztere proportional zu $p(x,y=x)$ ist.



Bei nichtlinearen Problemen kann $p(x|y_o)$ mehrere relative Maxima besitzen:



Diese "Mehrgipfligkeit" von $p(x|y_o)$ spiegelt die Mehrdeutigkeit des betreffenden Problems wider. Die Wdf $p(x|y_o)$ präsentiert den <u>vollständigen Wissensstand</u> nach dem Experiment. In der Praxis wird diese Information kondensiert durch Angabe der Lage der Maxima und ihrer "asymptotischen Varianz" (cf. Abschnitt 4).

3. Lineare Inversion mit Vorinformation

Im Falle einer linearen Theorie, d.h. für

$$\underline{y} = \underline{f}(\underline{x}) = \underline{G}\underline{x} \tag{3.1}$$

mit der (n x m)-Matrix $\underline{G}$ und Gaußscher Fehlerstatistik (2.7a,b) laßt sich die Bestimmung der a-posteriori Wdf $p(\underline{x}|\underline{y}_o)$ einfach geschlossen durchführen. Wie schon das simple Beispiel im vorangehenden Abschnitt andeutete, ergibt sich für $p(\underline{x}|\underline{y}_o)$ im linearen Fall stets eine Gaußverteilung, die durch Mittelwert $\hat{\underline{x}}$ und Kovarianzmatrix $\underline{C}_{\hat{x}}$ vollständig bestimmt ist. Deshalb sind nur diese Größen zu bestimmen.

Zwei Wege bieten sich an. Der erste geht von der Extremalbedingung (2.12) aus und liefert mit $\underline{f}(\hat{\underline{x}}) = \underline{G}\hat{\underline{x}}$ und $\underline{A}(\hat{\underline{x}}) = \underline{G}$

$$\underline{G}^T\underline{C}_y^{-1}(\underline{G}\hat{\underline{x}}-\underline{y}_o) + \underline{C}_x^{-1}(\hat{\underline{x}}-\underline{x}_o) = \underline{0},$$

so daß

$$\hat{\underline{x}} = (\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}(\underline{C}_x^{-1}\underline{x}_o+\underline{G}^T\underline{C}_y^{-1}\underline{y}_o) \tag{3.2}$$

$$= \underline{K}\underline{x}_o + \underline{L}\underline{y}_o \tag{3.3}$$

mit

$$\underline{K} = (\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}\underline{C}_x^{-1},$$

$$\underline{L} = (\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}\underline{G}^T\underline{C}_y^{-1}.$$

Wir wollen annehmen, daß für alle Komponenten von $\underline{x}$ endliche a-priori-Schätzwerte für die Streuung vorliegen. Dann ist $\underline{C}_x^{-1}$ positiv definit und die Inverse in (3.2) existiert für beliebiges m und n. Insbesondere sorgt auch in einem unterbestimmten System ($n < m$) die in $\underline{C}_x$ steckende Vorinformation dafür, daß eine eindeutige Lösung existiert und so die Inversion stabilisiert. Durch die Zerlegung (3.3) wird $\hat{\underline{x}}$ in seine beiden Anteile aus Vorinformation und Daten aufgespalten. Allerdings taucht noch in $\underline{K}$ und $\underline{L}$ die jeweils andere Kovarianzmatrix auf.

$\hat{\underline{x}}$ ist ein unverzerrter Schätzwert für das wahre aber unbekannte $\tilde{\underline{x}}$. Denn mit (2.6) und (3.3) ergibt sich

$$\langle \hat{\underline{x}} \rangle = \underline{K}\langle \underline{x}_o \rangle + \underline{L}\langle \underline{y}_o \rangle = \underline{K}\tilde{\underline{x}} + \underline{L}\tilde{\underline{y}} = (\underline{K} + \underline{L}\underline{G})\tilde{\underline{x}} = \tilde{\underline{x}} \tag{3.4}$$

Dabei bedeutet $\langle \cdot \rangle$ den statistischen Erwartungswert. Aus (2.6) folgt

$$\hat{\underline{x}} - \tilde{\underline{x}} = \underline{K}\underline{e}_x + \underline{L}\underline{e}_y,$$

so daß man als a-posteriori-Kovarianzmatrix erhält

$$\underline{C}_{\hat{x}} = \langle (\hat{\underline{x}} - \langle \hat{\underline{x}} \rangle)(\hat{\underline{x}} - \langle \hat{\underline{x}} \rangle)^T \rangle = \langle (\hat{\underline{x}} - \tilde{\underline{x}})(\hat{\underline{x}} - \tilde{\underline{x}})^T \rangle =$$

$$= \underline{K}\langle \underline{e}_x \underline{e}_x^T \rangle \underline{K}^T + \underline{L}\langle \underline{e}_y \underline{e}_y^T \rangle \underline{L}^T = \underline{K}\underline{C}_x\underline{K}^T + \underline{L}\underline{C}_y\underline{L}^T = (\underline{C}_x^{-1} + \underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}.$$

Zusammengefaßt:

$$\boxed{\underline{C}_{\hat{x}} = (\underline{C}_x^{-1} + \underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}} \tag{3.5}$$

$$\boxed{\hat{\underline{x}} = \underline{C}_{\hat{x}}(\underline{C}_x^{-1}\underline{x}_o + \underline{G}^T\underline{C}_y^{-1}\underline{y}_o)} \tag{3.6}$$

Vergleiche (3.5,6) mit (2.16)! Das Ergebnis (3.5,6) hätte man auch sehr einfach dadurch erhalten können, daß man die Koeffizienten von $\underline{x}^T$ und $\underline{x}^T\underline{x}$ in der Identität

$$(\underline{G}\underline{x} - \underline{y}_o)^T\underline{C}_y^{-1}(\underline{G}\underline{x} - \underline{y}_o) + (\underline{x} - \underline{x}_o)^T\underline{C}_x^{-1}(\underline{x} - \underline{x}_o) = (\underline{x} - \hat{\underline{x}})^T\underline{C}_{\hat{x}}^{-1}(\underline{x} - \hat{\underline{x}}) + \text{const.}$$

vergleicht.

Wie zu erwarten war, wird durch jede zusätzliche Messung der Wissensstand erweitert. Genauer: Für jede beliebige Linearkombination $\underline{b}^T\underline{x}$ der unbekannten Parameter (also insbesondere auch für jede Komponente von $\underline{x}$) ist die a-posteriori-Varianz nie größer als die a-priori Varianz, d.h.

$$\underline{b}^T\underline{C}_{\hat{x}}\underline{b} \leq \underline{b}^T\underline{C}_x\underline{b}. \tag{3.7}$$

Zum Beweis von (3.7) wird (3.5) umgeformt:

$$\underline{C}_{\hat{x}} = (\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}\left[(\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})\underline{C}_x - \underline{G}^T\underline{C}_y^{-1}\underline{G}\underline{C}_x\right] =$$

$$= \underline{C}_x - (\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}\underline{G}^T\underline{C}_y^{-1}\underline{G}\underline{C}_x =$$

$$= \underline{C}_x - \underline{C}_x\underline{G}^T(\underline{G}\underline{C}_x\underline{G}^T+\underline{C}_y)^{-1}\underline{G}\underline{C}_x \,. \qquad (3.8)$$

Dabei wurde von der Identität

$$(\underline{C}_x^{-1}+\underline{G}^T\underline{C}_y^{-1}\underline{G})^{-1}\underline{G}^T\underline{C}_y^{-1} = \underline{C}_x\underline{G}^T(\underline{G}\underline{C}_x\underline{G}^T+\underline{C}_y)^{-1}$$

Gebrauch gemacht, die man sofort durch Ausmultiplizieren verifiziert. Aus (3.8) folgt

$$\underline{b}^T\underline{C}_{\hat{x}}\underline{b} = \underline{b}^T\underline{C}_x\underline{b} - |(\underline{G}\underline{C}_x\underline{G}^T+\underline{C}_y)^{-1/2}\underline{G}\underline{C}_x\underline{b}|^2,$$

womit (3.7) bewiesen ist.

Diskussion von Grenzfällen:

Es sei $\underline{C}_x = \sigma_x^2\,\underline{I}_m$, $\underline{C}_y = \sigma_y^2\underline{I}_n$.

a) $\sigma_x \to 0$ oder $\sigma_y \to \infty$: Starke a-priori-Information oder ungewisse Daten

In diesem Grenzfall ist $\hat{\underline{x}}$ unabhängig von $\underline{y}_o$ und es gilt

$\hat{\underline{x}} = \underline{x}_o$, $\underline{C}_{\hat{x}} = \underline{C}_x$

b) $\sigma_x \to \infty$ oder $\sigma_y \to 0$: Schwache a-priori-Information oder sichere Daten

Jetzt läßt sich die Lösung durch die Pseudoinverse $\underline{G}^+$ von $\underline{G}$ ausdrücken und es gilt

$\hat{\underline{x}} = \underline{G}^+\underline{y}_o + (\underline{I}_m - \underline{G}^+\underline{G})\underline{x}_o$.

Im rein überbestimmten System (Rang $\underline{G}$ = m) ist $\underline{G}^+ = (\underline{G}^T\underline{G})^{-1}\underline{G}^T$ und man erhält die Lösung nach der Methode der kleinsten Quadrate

$\hat{\underline{x}} = (\underline{G}^T\underline{G})^{-1}\underline{G}^T\underline{y}_o$, $\underline{C}_{\hat{x}} = \sigma_y^2\,(\underline{G}^T\underline{G})^{-1}$.

Die Lösung ist unabhängig von $\underline{x}_o$.

Im rein unterbestimmten System (Rang $\underline{G}$ = n) macht sich die schwache Vorinformation noch bemerkbar und man erhält die

Lösung, die vom "a-priori-Punkt" $\underline{x}_o$ den kleinsten Abstand hat:

$$\hat{\underline{x}} = \left[\underline{I}_m - \underline{G}^T(\underline{G}\underline{G}^T)^{-1}\underline{G}\right]\underline{x}_o + \underline{G}^T(\underline{G}\underline{G}^T)^{-1}\underline{y}_o,$$

$$\underline{C}_{\hat{x}} = \left[\underline{I}_m - \underline{G}^T(\underline{G}\underline{G}^T)^{-1}\underline{G}\right]\sigma_x^2 + \underline{G}^T(\underline{G}\underline{G}^T)^{-2}\underline{G}\,\sigma_y^2.$$

Es folgen noch drei Anmerkungen zur oben gegebenen Lösung für das lineare Problem:

1) Randverteilungen

Die a-posteriori-Wdf ist

$$p(\underline{x}|\underline{y}_o) = \frac{1}{\left[(2\pi)^m \det \underline{C}_{\hat{x}}\right]^{1/2}} \exp\left\{-\tfrac{1}{2}(\underline{x} - \hat{\underline{x}})^T \underline{C}_{\hat{x}}^{-1}(\underline{x} - \hat{\underline{x}})\right\} .$$

Wenn man sich nur für die Verteilung eines Parameters, etwa x_1, interessiert, betrachtet man die zu x_1 gehörende Randverteilung, die man durch Integration über die übrigen Variablen erhält:

$$p(x_1|\underline{y}_o) = \int p(\underline{x}|\underline{y}_o)\, dx_2 \ldots dx_m.$$

Entsprechendes gilt für mehrere interessierende Parameter.

2) Relative Bedeutung von Vorinformation und Daten

Bei Verwendung von Vorinformation (d.h. $\underline{C}_x$ nicht singulär) werden stets alle m Parameter aufgelöst. Wegen (3.4), d.h.

$$\langle \hat{\underline{x}} \rangle = (\underline{K} + \underline{L}\underline{G})\,\tilde{\underline{x}} = \underline{I}_m\tilde{\underline{x}} = \tilde{\underline{x}}$$

ist die Summe der Diagonalelemente von $\underline{K}$ und $\underline{L}\underline{G}$ gleich der Zahl m der Parameter. Die Summe der Diagonalelemente von $\underline{K}$ bzw. $\underline{L}\underline{G}$ vermittelt eine Vorstellung davon, wieviele Parameter durch Vorinformation bzw. Daten aufgelöst werden.

3) Fehlerhafte oder ungenau bekannte Theorie

Wenn die "Theorie" $\underline{y} = \underline{f}(\underline{x})$ entweder nicht genau bekannt ist oder zur Vereinfachung der Interpretation absichtlich eine ungenaue Theorie benutzt wird, kann man versuchsweise $\underline{y} - \underline{f}(\underline{x})$ als einen Zufallsvektor mit bekannter Statistik betrachten. Im einfachsten Fall einer Normalverteilung lautet etwa die n-dimensionale bedingte Wdf

$$p(\underline{y}|\underline{x}) = \left[(2\pi)^n \det \underline{C}_T\right]^{-1/2} \exp\left\{-\tfrac{1}{2}(\underline{y}-\underline{f}(\underline{x}))^T \underline{C}_T^{-1}(\underline{y}-\underline{f}(\underline{x}))\right\}$$

mit vorgegebener Kovarianzmatrix $\underline{C}_T$. Im Fall einer exakten Theorie ist $p(\underline{y}|\underline{x}) = \delta(\underline{y} - \underline{f}(\underline{x}))$.

Im linearen Fall, d.h. für

$$\underline{y} = \underline{G}\underline{x} + \underline{e} \text{ mit } \langle \underline{e}\underline{e}^T \rangle = \underline{C}_T$$

anstelle von $\underline{y} = \underline{G}\underline{x}$ führt die fehlerhafte Theorie lediglich dazu, daß in (3.5) und (3.6) $\underline{C}_y$ durch $\underline{C}_y + \underline{C}_T$ ersetzt wird, d.h. Datenrauschen und "geologisches Rauschen" addieren sich einfach.

4. Nichtlineare Inversion mit Vorinformation

Wenn $\underline{y} = \underline{f}(\underline{x})$ nichtlinear ist, muß zur Bestimmung von $\underline{x}$ das System (2.12) i.a. iterativ gelöst werden. Auch bei Gaußscher Fehlerstatistik können sich mehrere Lösungen ergeben (siehe das Beispiel in Abschnitt 2). Die Lösungen bezeichnen relative Maxima der der a-posteriori-Wdf. Das größte Maximum ist das globale Maximum. Die Wdf $p(\underline{x}|\underline{y}_o)$,die nun keine Gaußverteilung mehr ist, enthält die vollständige Information über unseren Wissensstand und ist das eigentliche Ergebnis der Inversion. Diese m-dimensionale Verteilung ist sehr unhandlich und muß kondensiert werden.

Die naheliegendste Möglichkeit besteht darin, in der Umgebung eines Maximums $\hat{\underline{x}}$ zu linearisieren und so die zu $\underline{x}$ gehörende Kovarianzmatrix $\underline{C}_x^a$ zu bestimmen, die nur im Falle geringer Nichtlinearität für $p(\underline{x}|\underline{y}_o)$ repräsentativ ist und deshalb als asymptotische Kovarianzmatrix bezeichnet wird. In Analogie zu (3.5) ist $\underline{C}_x^a$ definiert durch

$$\underline{C}_{\hat{x}}^a = \left[\underline{C}_x^{-1} + \underline{A}(\hat{\underline{x}})^T \underline{C}_y^{-1} \underline{A}(\hat{\underline{x}})\right]^{-1} \tag{4.1}$$

Die eigentlichen Kovarianzen und Erwartungswerte der Parameter bestimmen sich aus der tatsächlichen a-posteriori-Wdf:

$$\hat{\underline{\hat{x}}} = \int \underline{x}\, p(\underline{x}|\underline{y}_o)\, d\underline{x}\,, \tag{4.2}$$

$$\underline{C}_{\hat{x}} = \int (\underline{x} - \hat{\underline{\hat{x}}})(\underline{x} - \hat{\underline{\hat{x}}})^T p(\underline{x}|\underline{y}_o)\, d\underline{x}\,. \tag{4.3}$$

Im linearen Fall stimmen die nach (4.2) und (4.3) berechneten Größen mit $\underline{\hat{x}}$ und $\underline{C}_{\hat{x}}$ nach (3.6) und (3.5) überein. Im nichtlinearen Fall braucht natürlich dem $\underline{\hat{\hat{x}}}$ nach (4.2) kein Maximum von $p(\underline{x}|\underline{y}_o)$ zu entsprechen. Die Berechnung von $\underline{\hat{\hat{x}}}$ und $\underline{C}_{\hat{\hat{x}}}$ nach (4.2,3) ist aufwendig, da jeweils m-dimensionale Integrale ausgewertet werden müssen.

Zur Lösung des nichtlinearen Systems (2.12) schlagen Jackson & Matsu'ura (1985) den folgenden Iterationsalgorithmus vor:

$$\underline{x}_{k+1} = \underline{x}_k + \mu \underline{C}_k \left\{ \underline{C}_x^{-1} (\underline{x}_o - \underline{x}_k) + \underline{A}_k^T \underline{C}_y^{-1} (\underline{y}_o - \underline{f}_k) \right\} \qquad (4.4)$$

mit $\underline{A}_k := \underline{A}(\underline{x}_k)$, $\underline{f}_k := \underline{f}(\underline{x}_k)$,

$$\underline{C}_k := (\underline{C}_x^{-1} + \underline{A}_k^T \underline{C}_y^{-1} \underline{A}_k)^{-1},$$

$0 < \mu \leq 1$.

Als Startvektor kann das a-priori-Modell $\underline{x}_o$ dienen. Der die Schrittweite begrenzende Parameter μ sorgt für eine Stabilisierung. Für schwach nichtlineare Probleme kann $\mu = 1$ gewählt werden. Im linearen Fall führt dann (4.4) für jeden beliebigen Startvektor in einem Schritt zum Ziel (3.6).

Für $\mu = 1$ ist äquivalent zu (4.4)

$$\underline{x}_{k+1} = \underline{x}_o + \underline{C}_k \underline{A}_k^T \underline{C}_y^{-1} \left\{ \underline{y}_o - \underline{f}_k - \underline{A}_k (\underline{x}_o - \underline{x}_k) \right\}, \qquad (4.5)$$

wie man sofort durch Betrachtung der Differenz (4.5) - (4.4) erkennt. Jeder Iterationsschritt benötigt zur Berechnung von $\underline{C}_k$ die Inversion einer (m x m)-Matrix. Für unterbestimmte Systeme ($n < m$) wird deshalb (4.5) modifiziert mit Hilfe der Identität

$$\underline{C}_k \underline{A}_k^T \underline{C}_y^{-1} = \underline{C}_x \underline{A}_k^T (\underline{C}_y + \underline{A}_k^T \underline{C}_y \underline{A}_k)^{-1}.$$

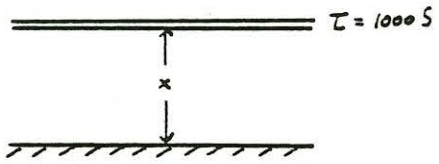
Eine Stabilisierung von (4.5) kann in Analogie zur Marquardt-Methode dadurch erreicht werden, daß man in $\underline{C}_k$ die Größe $\underline{C}_x^{-1}$ durch $\underline{C}_x^{-1} + \mu \underline{I}_m$ ersetzt, wobei μ eine geeignet gewählte positive Zahl ist. Dies entspricht einer Verstärkung der Vorinformation.

5. Ein einfaches Beispiel

In einem Problem der elektromagnetischen Tiefensondierung sei für die Periode $T = 2\pi/\omega = 1800$ s die Übertragungsfunktion

$$c = \frac{E_x}{i\omega\mu_o H_y} = y_{o1} - iy_{o2} = (100 - i100)\ \text{km} \qquad (5.1)$$

gegeben. Die Standardabweichungen seien $\sigma_{y1} = \sigma_{y2} =: \sigma_y = 20$ km. Diese beiden Daten mögen interpretiert werden durch eine dünne Deckschicht mit dem bekannten Leitwert $\tau = 1000$ S und einen idealen Leiter in der unbekannten Tiefe x. Für x sei jedoch die Vorinformation $x_o = 250$ km mit der Standardabweichung $\sigma_x = 50$ km bekannt. Wie sieht der Wissensstand nach Berücksichtigung der obigen Meßdaten aus ?



Im vorliegenden Problem ist $m = 1$, $n = 2$. Der theoretische Zusammenhang zwischen c, τ, T und x ist

$$c = \frac{x}{1 + i\omega\mu_o\tau x}, \quad \omega = 2\pi/T,$$

so daß nach (5.1) mit $\beta := \omega\mu_o\tau = (228\ \text{km})^{-1}$ gilt:

$$\underline{f}(x) = \frac{x}{1 + \beta^2x^2}\begin{pmatrix}1\\ \beta x\end{pmatrix},$$

$$\underline{A}(x) = \frac{1}{(1 + \beta^2x^2)^2}\begin{pmatrix}1 - \beta^2x^2\\ 2\beta x\end{pmatrix}.$$

Aus (2.12) ergibt sich damit nach einfacher Zwischenrechnung als Bestimmungsgleichung für $\hat{x}$:

$$\hat{x} - (1-\beta^2\hat{x}^2)y_{o1} - 2\beta\hat{x}y_{o2} + (\sigma_y/\sigma_x)^2(1+\beta^2\hat{x}^2)^2(\hat{x}-x_o) = 0.$$

Diese algebraische Gleichung 5. Ordnung besitzt nur die eine reelle Lösung $\hat{x} = 218.4$. Ohne Vorinformation ($\sigma_x = \infty$) hätte man $x_\infty = 198$ km erhalten.

Die asymptotische Varianz nach (4.1) beträgt

$$(C_x^a)^{-1} = \frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2(1 + \beta^2 x^2)}$$

so daß $\sigma_x^a = 24.2$ km.

Mittelwert und Streuung der a-posteriori - Verteilung $p(x|\underline{y}_o)$ erhält man nach (4.2,3) durch Berechnung der drei Momente

$$M_k := \int_0^\infty x^k \exp\{-\tfrac{1}{2} F(x)\}\, dx, \quad k = 0, 1, 2$$

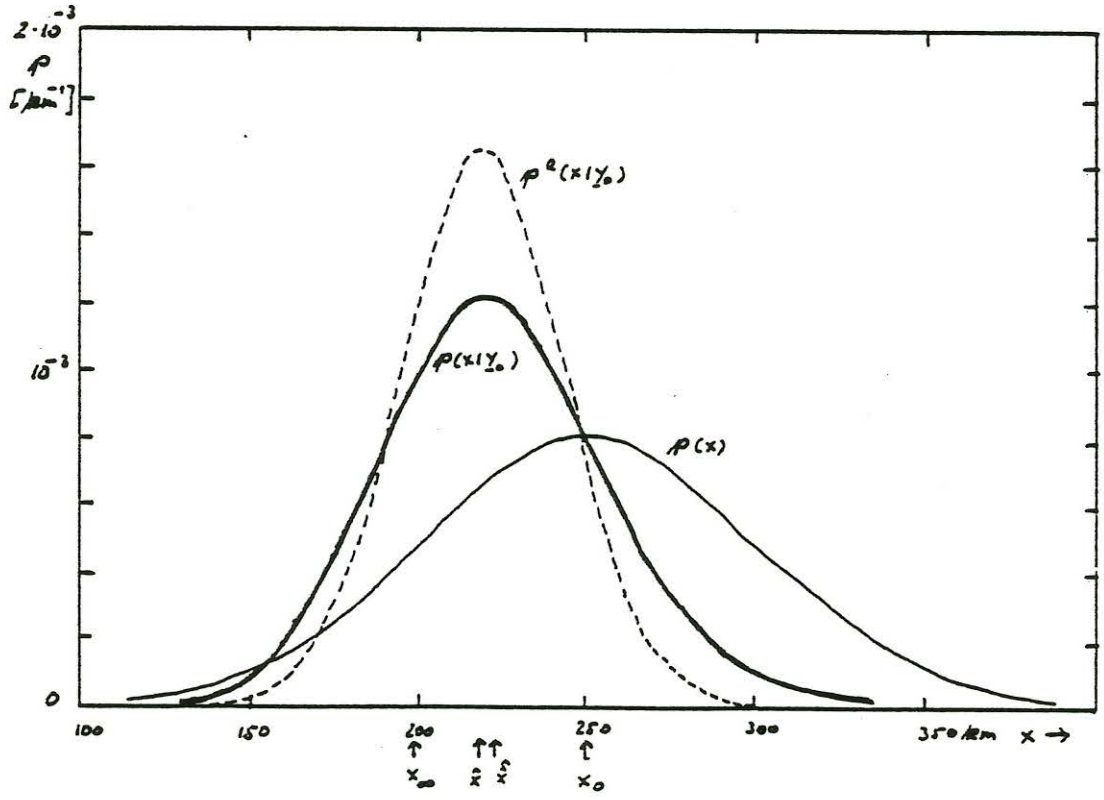
mit

$$F(x) = \left(\frac{x-x_o}{\sigma_x}\right)^2 + \left\{\left(\frac{x}{1+\beta^2x^2} - y_{o1}\right)^2 + \left(\frac{\beta x^2}{1+\beta^2x^2} - y_{o2}\right)^2\right\}/\sigma_y^2 .$$

Dabei dient M_o zur Normierung von $p(x|\underline{y}_o)$. Gl. (4.2,3) liefert dann

$$\hat{\hat{x}} = M_1/M_o = 222.8 \text{ km}, \quad \sigma_{\hat{\hat{x}}} = (M_o M_2 - M_1^2)^{1/2}/M_o = 33.0 \text{ km}.$$

Die tatsächliche Standardabweichung $\sigma_{\hat{\hat{x}}}$ ist also größer als die asymptotische Standardabweichung σ_x^a. Die folgende Skizze zeigt die a-priori-Wdf $p(x)$, die a-posteriori-Wdf $p(x|\underline{y}_o)$ sowie ihre Approximation durch die asymptotische Verteilung $p^a(x|\underline{y}_o)$ (Normalverteilung).



6. Literaturempfehlungen

Die folgenden Arbeiten wurden entweder im Text erwähnt (1 - 5) oder enthalten wichtige Beiträge zum Begriff der Vorinformation (6 - 7). In den Arbeiten 1,2,4,5 wird die Theorie dargestellt und mit zahlreichen Beispielen aus der Geophysik illustriert. Dabei behandeln 4 und 5 auch kontinuierliche Modelle.

1) Jackson, D.D. & Matsu'ura, M.: A Bayesian approach to nonlinear inversion. J. Geophys. Res., 90, 581-591, 1985

2) Menke, W.: Geophysical data analysis: Discrete inverse theory. Academic Press 1984

3) Parker, R.L.: The inverse problem of electromagnetic induction: existence and construction of solutions based on incomplete data. J. Geophys. Res., 85, 4421-4428, 1980

4) Tarantola, A. & Valette, B.: Inverse problems = Quest for information. J. Geophys., 50, 159-170, 1982a

5) Tarantola, A. & Valette, B.: Generalized nonlinear inverse problems solved using the least squares criterion. Rev. Geophys. Space Phys., 20, 219-232, 1982b

6) Jaynes, E.T.: Prior probabilities. IEEE Trans. Systems, Science and Cybernetics, 4, 227-241, 1968

7) Rietsch, E.: The maximum entropy approach to inverse problems. J. Geophys., 42, 489-506, 1977