Chapter

Introductory Chapter: Ramifications of Incomplete Knowledge

Jan Peter Hessling

"*Facts do not cease to exist because they are ignored*."

1. Background

Mathematical statistics has long been widely practiced in many fields of science [1]. Nevertheless, statistical methods have remained remarkably intact ever since the pioneering work [2] of R.A. Fisher and his contemporary scientists early in the twentieth century. Recently however, it has been claimed that most scientific results are wrong [3], due to malpractice of statistical methods. Errors of that kind are not caused by imperfect methodology but rather, reflect lack of understanding and proper interpretation.

In this introductory chapter, a different cause of errors is addressed—the ubiquitous practice of willful ignorance (WI) [4]. Usually it is applied with intent to remedy lack of knowledge and simplify or merely enable application of established statistical methods. Virtually all statistical approaches require complete statistical knowledge at some stage. In practice though, that can hardly ever be established. For instance, Bayes estimation relies upon prior knowledge. Any equal a priori probability assumption ("uninformed prior") does hardly disguise some facts are not known, which may be grossly deceiving. Uniform distribution is a specific assumption like any other. Willful ignorance of that kind must not be confused with knowledge to which we associate some degree of confidence. It may be better to explore rather than ignore consequences of what is not known at all. That will require novel perspectives on how mathematical statistics is practiced, which is the scope of this book.

2. Ambiguity

Incomplete knowledge implies that obtained results may not be unique. That is, results may be ambiguous. Ambiguity de facto means the uncertainty associated with any estimated quantity itself is uncertain. We may adopt a probabilistic view and classify ambiguity as epistemic uncertainty. Ambiguity will here refer to lack of knowledge typically substituted with willful ignorance. Alternatives propelled by different types of willful ignorance can thus be explored to assess ambiguity.

A most powerful source of ambiguity is dependencies. Independence is perhaps the most claimed but often the least discussed presumption. Throwing dices or growing crops, as typically studied by the founders of statistics, independence indeed seems plausible. In all the complexity of modern technology of today however, it is anything but evident observations are independent. For instance,

meteorological radar observations may share sources of errors, meaning recorded data will be statistically dependent. A problem may then arise if our analysis makes use of, e.g., the maximum likelihood method which utilizes the entire covariance matrix. Most of its entries, all covariances between pairs of observations, are usually not known but bluntly set to zero to enable evaluation. This willful ignorance has the drastic consequence of extinguishing ambiguity and, as will be shown, minimizing the resulting uncertainty. Elementary considerations should provide the valuable insight that even exceedingly small covariances may substantially influence the result: the number of covariance elements is $n(n-1)/2 \approx n^2/2$, while there are only n variances, for n observations. The number of covariance elements is hence $n/2$ times larger than that of variance. Each element being small is thus not a good enough argument to ignore the collection of all covariance elements.

Various attempts have been made to avoid willful ignorance. The method of maximum entropy [5] focuses on the consequences of improper assignments of unknown statistical information. Covariance intersection [6] fuses observations conservatively to a pair of uncorrelated observations with variance $\max[\mathrm{var}(\theta)]$. This approach explores ambiguity along the general principles suggested here, considering all possible values of covariance. Complementing the obtained maximum variance with the least possible variance $\min[\mathrm{var}(\theta)]$ would render an ambiguity interval, $\mathrm{A} = \max[\mathrm{var}(\theta)] - \min[\mathrm{var}(\theta)]$, different but similar to confidence intervals.

Repeating any statistical analysis with various kinds of willful ignorance [on its input], the ambiguity (A) [of its output] can be assessed. Some WI will give large, while others will yield small resulting uncertainty, not necessarily the maximum and minimum, as it is difficult to imagine all possible kinds of WI. Any specific WI will more or less reduce or quench the uncertainty from its maximum. Identifying a model from calibration data H_{CAL} and then letting the so-obtained model predict the same data H_{PRD}, any chosen willful ignorance of $\mathrm{cov}(H_{\mathrm{CAL}})$ will quench the calibration uncertainty from the maximum over all choices, $\mathrm{var}(H_{\mathrm{PRD}}) \leq \max[\mathrm{var}(H_{\mathrm{PRD}})] \leq \mathrm{var}(H_{\mathrm{CAL}})$. Studying uncertainty quenching through $\mathrm{var}(H_{\mathrm{CAL}}) - \mathrm{var}(H_{\mathrm{PRD}})$ will indicate possible ramifications of our lack of knowledge $\mathrm{var}(H_{\mathrm{PRD}}) \leq \max[\mathrm{var}(H_{\mathrm{PRD}})]$ but also the implicit knowledge $\max[\mathrm{var}(H_{\mathrm{PRD}})] \leq \mathrm{var}(H_{\mathrm{CAL}})$ contained in the structure of the model. Most importantly, such studies will guide us to the least harmful choice of willful ignorance. The analysis is similar in style but different to the method of maximum entropy and covariance intersection. An example is given below.

3. Illustration of uncertainty quenching

Assume we would like to study the evolution of a field over two spatial coordinates, using a model composed of a set of differential equations. The field could refer to meteorology and describe current observations of air pressure or humidity. The initial state may be expanded in the set of basis functions of the appropriate operator, similar to forecasting in numerical weather prediction (NWP) [7]. The basis functions could be thought of as the eigensolutions of a linear operator, which propagates one meteorological state, from one day to another. Neither the interpretation of the field nor the field itself matters for the discussion here. Rather, it is how the uncertainty of the initial state is represented as uncertainty of the distributed eigensolutions of the NWP propagator. This representation will determine the uncertainty of any subsequent forecast, reflecting the past experience in future confidence of predicting the weather. If the forecast uncertainty is lower than our current knowledge reflects, we may falsely reject, e.g., the possibility of experiencing major thunderstorms. In the eye of sailors planning their journey, the forecast

uncertainty is the indisputable decision-maker. Studying the uncertainty quenching $\mathrm{var}(H_{\mathrm{CAL}}) - \mathrm{var}(H_{\mathrm{NWP}})$, the ambiguity regarding the usually unknown but nevertheless required covariances $\mathrm{cov}\left(H_{\mathrm{CAL}}(x_i, y_i), H_{\mathrm{CAL}}\left(x_j, y_j\right)\right)$ can be assessed. Then by expanding the uncertainty conservatively, serious events like major thunderstorms may be properly recovered.

To enable illustrations, let the eigenstates of the NWP operator of order n [for not known reasons] be multiplicative separable in time t as well as in spatial coordinates x, y, with eigenstates described by orthogonal Legendre polynomials up to order n:

$$H_{\mathrm{NWP}}^{(n)}(x,y,t) \equiv \sum_{j,\,k=0}^{n} \underbrace{\theta_{j+k(n+1)+1}}_{\equiv r}(t) \cdot P_k(y) P_j(x), \tag{1}$$

where the NWP operator propagates the coefficients $\theta_{j+k(n+1)+1}(t)$ in time. Only the representation of $\mathrm{cov}(H_{\mathrm{CAL}}(x,y,t=0))$, the covariance of the measured initial state at $t = 0$, is of interest here. Discretizing over m domains in both directions x, y followed by sequential scanning over x_p for each y_q, $p, q = 1, 2, \ldots m+1$, the model is written in standard affine vector $\left(\overline{\theta}, \overline{K}\right)$ form:

$$H_{\mathrm{NWP}}^{(n)}\left(x_p, y_q, 0\right) \equiv \sum_{r=1}^{(n+1)^2} \theta_r(t=0) K_r\left(x_p + (m+1) y_q\right) = \overline{\theta}^T \overline{K}. \tag{2}$$

Without any supplementary information, the variance of the initial measurement should be completely represented by the variance of the initial model state, i.e., $\mathrm{var}(H_{\mathrm{NWP}}(x,y,0)) = \mathrm{var}(H_{\mathrm{CAL}}(x,y,0))$. The question is whether this holds, and if it does not, to which extent can we minimize the discrepancy with WI?

Assuming normal distributed measurement noise, the maximum likelihood method [8] yields the parameter covariance given by Eq. (3), which is propagated to uncertainty of the best predictions according to Eq. (4):

$$\mathrm{cov}(\theta^*) = \left[K \mathrm{cov}(H_{\mathrm{CAL}}) K^T\right]^{-1}, \tag{3}$$

$$\mathrm{cov}(H_{\mathrm{NWP}}) = K^T \mathrm{cov}(\theta^*) K \tag{4}$$

Combining these relations, the degree of completeness of the representation of uncertainty by the model can be studied:

$$\mathrm{var}(H_{\mathrm{NWP}}) = \mathrm{diag}\left\{K^T \left[K \mathrm{cov}(H_{\mathrm{CAL}}) K^T\right]^{-1} K\right\} \cong \mathrm{var}(H_{\mathrm{CAL}}), \tag{5}$$

where $\cong$ indicates the addressed equality in the absence of uncertainty quenching or maximal propagation of uncertainty from observations to model. Equality can never be achieved though, since the number of degrees of freedom (NDOF) of prediction is drastically lower than that of calibration data. For typical models and data, the two NDOFs usually differ by an order of 10 or more. A large ratio is actually required to provide sufficient redundancy. As seen in **Figure 1**, the uncertainty is normally reduced to a small fraction, with substantial uncertainty quenching.

It should be emphasized that stating independence is fundamentally different than stating the degree of dependence which is unknown. These statements in fact oppose each other, since independence maximizes the available amount of information. Indeed, the Fisher information matrix [9]

$$\frac{1}{\mathrm{CRLB}} = F\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right) = -\left\langle \frac{\partial^2 \ln(p)}{\partial H_{\mathrm{CAL},i} \partial H_{\mathrm{CAL},j}} \right\rangle = \begin{cases} F_1, & m \text{ dep.} \\ mF_1, & m \text{ indep.} \end{cases} \tag{6}$$

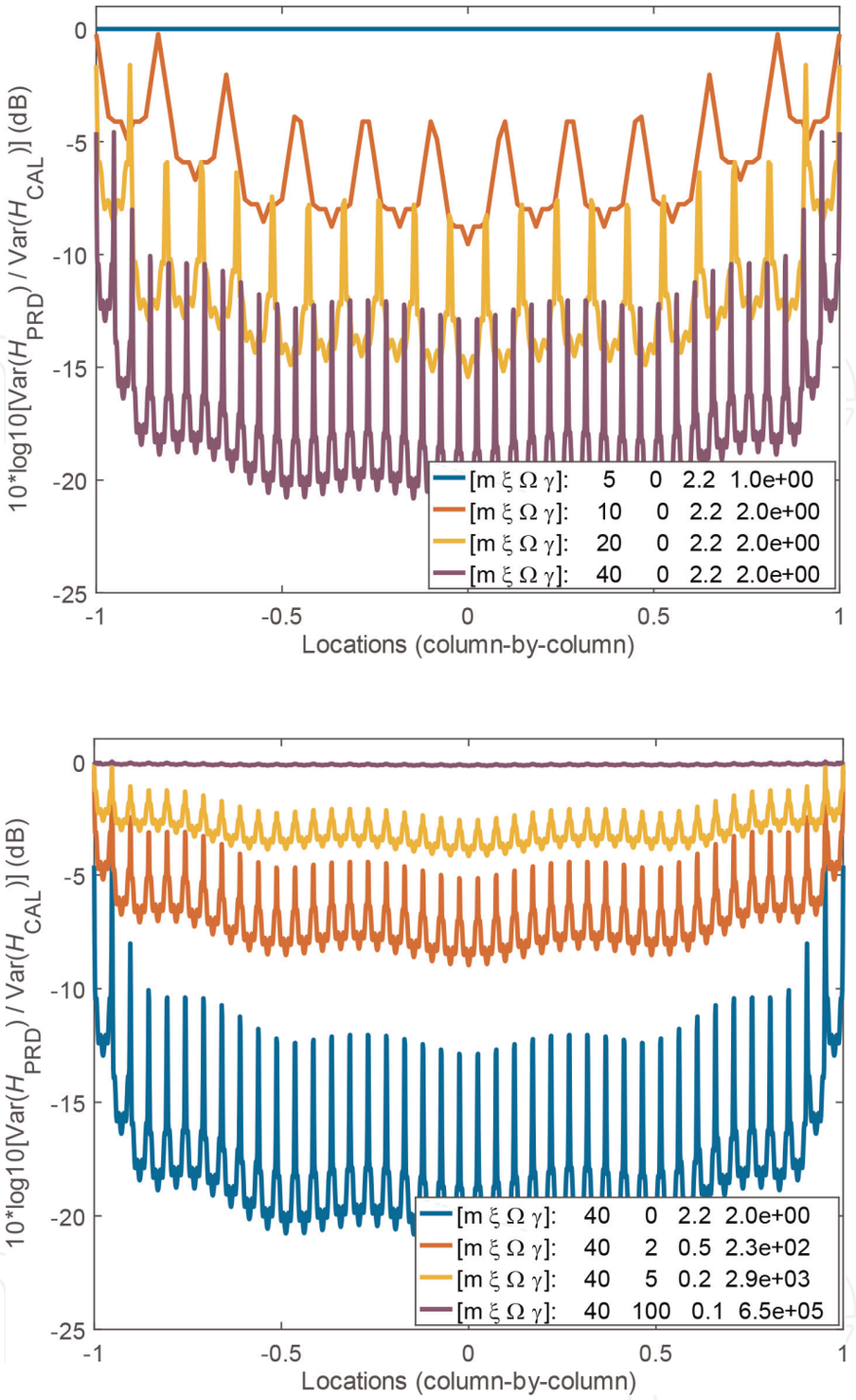

Figure 1.
Uncertainty quenching or excessive reduction of uncertainty due to willful ignorance. Dependence on resolution m (top) and correlation length ξ (Eq. (7)) (bottom) of the calibration data. The legend includes the Mahalanobis canonical distance Ω (Eq. (11)) and the ratio γ between the largest and the smallest eigenvalues of residual variance (Eq. (9)).

is additive as m independent observations are collected, since the joint probability distribution p in that case is multiplicative separable. For dependent observations though, no information is added at all, since $F(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j})$ then remains the same as for only one observation (F_1). Aggregating m observations, the minimum variance of any estimator set by the Cramer-Rao lower bound (CRLB) [9] thus decreases with a factor $1/m$ in the case where errors are independent (Eq. (6), first row), while it remains the same if they are completely dependent (Eq. (6), second row). Independence is thus the worst possible choice of WI, as it builds

confidence without knowledge. WI should minimize rather than maximize the information. That is indeed the principle utilized in the method of maximum entropy [5].

Uncertainty is lost for obvious reasons. The question is how much and for what reason. Since the model cannot represent an arbitrary response, it can neither represent an arbitrary variability. This restriction constitutes the very meaning of a "model." This makes it important to describe the covariance of observations accurately—inappropriate WI may quench uncertainty dramatically.

The additional information represented by the structure of the model could be denoted by the model innovation. It is strongly affected by WI attributed to observations. With increasing resolution m, the model innovation grows as the information contained in observations is maximized with an assumption of independence. Indeed, the prediction variance is quenched in agreement with the CRLB, as seen in **Figure 1** (top).

If WI of observation covariance instead resembles what the model is able to represent, the model innovation will be the least. Instead of assuming independent observations, introduce a finite long correlation length\ksi:

$$\operatorname{corr}\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right) = \frac{\operatorname{cov}\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right)}{\sqrt{\operatorname{var}(H_{\mathrm{CAL},i})\operatorname{var}\left(H_{\mathrm{CAL},j}\right)}} = \exp\left[-\left|r_i - r_j\right|/\xi\right] \tag{7}$$

Increasing the correlation length\ksi from zero as in **Figure 1** (bottom), the model innovation decreases, and the variance of the prediction $\operatorname{var}(H_{\mathrm{NWP}})$ is almost fully restored to the original variance of the observation $\operatorname{var}(H_{\mathrm{CAL}})$. The model will then not improve our knowledge of the current weather situation but enable prediction to a later time with comparable trust. Summarizing **Figure 1**, our WI of observation covariance and resolution m strongly influence our claimed precision $\operatorname{var}(H_{\mathrm{NWP}})$ of predictions.

It is a different matter if the model is consistent with the observations it was identified from. Model consistency is usually assessed with a statistical residual analysis. In conventional system identification (CSI) [10], the hypothesis is that the [deterministic] model fully explains the observations. Due to sampling variance of the finite uncertain calibration data though, the best estimate of its parameters will be uncertain. The residual analysis explores if the residual is consistent with the sampling uncertainty of the calibration data but without uncertainty associated with the model.

This conjecture of a model without error whatsoever in CSI is questionable. In practice, no model is completely without error. Rather, a finite uncertainty of the model could be regarded as inherited from mismatch to calibration data. If so, the model merely provides a convenient but to a quantifiable degree imperfect basis for expressing uncertain calibration data. The model is utilized to "passively transform" rather than "actively explain" observations to another unknown situation of interest. That intent is typical in, e.g., weather forecasting and product development. Furthermore, the uncertainty of calibration data can often be assessed from the setup of the calibration experiment. In CSI correlation functions are evaluated from a single residual vector, enforcing homoscedasticity and independence of observations. WI of this kind enables the statistical analysis of the residual but often find little support.

The alternative view on model calibration proposed here is that the identified model, composed of its form or structure, parameters, and uncertainty, represents the uncertain calibration data. Model results can thus substitute our observations, to the degree various aspects of the model and observations are consistent. Any given

residual is one realization and should relate to its expected variability, with respect to the uncertainties of both the model and the observations it was identified from.

The Mahalanobis distance [6] can be utilized to measure the relative distance between observations and model output, which constitutes the residual ρ:

$$M = \rho^T \mathrm{cov}^{-1}(\rho)\rho, \quad \rho \equiv H_{\mathrm{CAL}} - H_{\mathrm{PRD}}. \tag{8}$$

The residual covariance matrix defines its principal variations with typical magnitudes λ_j:

$$\begin{aligned} &\mathrm{cov}(\rho) = \mathrm{cov}(H_{\mathrm{CAL}}) - 2\mathrm{cov}(H_{\mathrm{CAL}}, H_{\mathrm{PRD}}) + \mathrm{cov}(H_{\mathrm{PRD}}) = U\Lambda^2 U^T, \\ &\Lambda_{ij} \equiv \delta_{ij}\lambda_j, \quad UU^T = I, \quad \delta_{jj} = 1, \quad \delta_{i\neq j} = 0. \end{aligned} \tag{9}$$

The evaluation of $\mathrm{cov}(H_{\mathrm{CAL}}, H_{\mathrm{PRD}})$ in Eq. (9) is challenging, since H_{PRD} has a complicated relation to its "role model" H_{CAL} set by the identification. To simplify, it is set to zero below.

Extracting matrices U from Eq. (9) in Eq. (8), squared deviations are compared to variances. The Mahalanobis distance then transforms into a relative Euclidean norm of the residual in its own space of uncorrelated variations:

$$M \equiv \tilde{\rho} T \Lambda^{-2} \tilde{\rho} = \sum_j \frac{\left|\tilde{\rho}_j\right|^2}{\mathrm{var}\left(\tilde{\rho}_j\right)}, \quad \tilde{\rho} \equiv U\rho, \tag{10}$$

where $\tilde{\rho}_j$ is the projection of the residual on its principal vector $U_{:,j}$ of variation, while the eigenvalue $\lambda_j \equiv \{\Lambda\}_{jj}$ expresses its typical magnitude of variation. For a small eigenvalue λ_j, observing even a moderate projection $U_{j,:}\rho$ is statistically unlikely and thus strongly violates any model.

To maximize the consistency, in the sense of minimizing the Mahalanobis distance, the variance $\mathrm{var}\left(\tilde{\rho}_j\right)$ of principal residual variations should be maximized. Without addressing any specific residual, maximize what could be defined the Mahalanobis canonical distance:

$$\Omega \equiv \sqrt{\sum_j \mathrm{var}\left(\tilde{\rho}_j\right)} = \sqrt{\sum_j \lambda_j^2}. \tag{11}$$

Minimizing the Fisher information matrix under assumption of normality addresses the covariance $\mathrm{cov}\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right)$ of observations. Minimizing the Mahalanobis canonical distance Ω considers also the covariance $\mathrm{cov}\left(H_{\mathrm{PRD},i}, H_{\mathrm{PRD},j}\right)$ of the model residual as well as the cross covariance $\mathrm{cov}\left(H_{\mathrm{CAL},i}, H_{\mathrm{PRD},j}\right)$, which reflects the model innovation. Hence, willful ignorance for model identification should minimize the Mahalanobis canonical distance rather than the Fisher information matrix, as the former but not the latter also accounts for the innovation of the model structure. The intent is to educate the model to produce the most conservative results.

In practice, no residual projection $\tilde{\rho} = U\rho$ is usually negligible. Thus, the likelihood of rejecting any model, considering correlations, increases dramatically if exceedingly small eigenvalues are obtained. For that reason it is wise to check the ratio $\gamma \equiv \max_j\left(\lambda_j\right) / \min_j\left(\lambda_j\right)$ between the largest and smallest eigenvalues of the residual covariance (Eq. (9)). If γ is large, the model is expected to fail with respect

to correlations of the residual. The model may very well be consistent with respect to the variance but rarely with respect to covariance of its output. However, ignoring correlations and only focusing on the magnitude of variations of calibration data, i.e., $\mathrm{var}(H_{\mathrm{CAL}})$, which is the standard practice [10], is completely different. Then, the belief in the model is perfect and the only limitation of also making perfect forecasts is the finiteness of a random sample of observations. In case of homoscedasticity, $\gamma = 1$.

A potential conflict is inevitable for exceedingly high ratios γ. Indeed, as seen in **Figure 1** (bottom) for increasing correlation lengths ξ, the Mahalanobis distance Ω decreases, while the ratio γ rapidly increases. Thus as observation variance is recovered, the requirement to ignore prediction covariance rapidly grows.

4. A quest for better practice of willful ignorance

> "*The first principle is that you must not fool yourself and you are the easiest person to fool*" [11].

Current practice of willful ignorance sometimes makes statistics an art of self-delusion [3]. Consequences of applied WI are rarely explored, as only one proposition normally is made without further ado.

Distinguishing what is not known from what is assumed is of paramount importance. Not known to any degree should mean that all possibilities that can be imagined also ought to be considered. Otherwise obtained results only exemplify what the most appropriate answer may be, without any indication of the largest possible deviation.

Our knowledge is almost never complete. Virtually all existing statistical methods nevertheless require precisely that. Until alternative methodologies exist, WI must fill the gap between what is actually known and what must be known. As illustrated, the consequences of different WI may vary dramatically. Therefore we should select and tweak WI carefully. WI should not relate to our unconfirmed belief, but rather address its consequences.

The proposal of a quantifiable ambiguity proposed here suggests how ramifications of incomplete knowledge might be mitigated with carefully chosen WI: explore all kinds of ignorance that can be imagined. Analyze and collect obtained results in ambiguity intervals, similar to confidence intervals. Another option is to focus on the worst case in a conservative manner. The method of covariance intersection is one example of how that can be exercised. The principle of maximum entropy provides means to maximize the residual uncertainty, to add the least possible amount of information. Minimizing the Fisher information for observations and the Mahalanobis distance for model identification as proposed here is still another kind of conservatism. These methods tackle unknown information with WI and explore its consequences. Finding the most proper WI is indeed nontrivial and calls for genuinely novel approaches.

Current practice of statistics utilizes WI in many ways, but the specific choice is rarely discussed in depth. One reason could be that statistics was developed in an entirely different context than practiced today, which is rarely acknowledged and probably not fully comprehended. To exemplify, recall that Fisher's [2] original interpretation of "never" as a finite probability of 5% was just a humble proposal. He urged his readers to adjust "never" to the current context, a piece of advice almost never followed today.

Perhaps the reported breakdown of statistics methodologies [3, 4] is due to neglect of ambiguity, driven by a strong tradition of uncritical application of WI.

Could this be caused by lack of awareness of its potentially dramatic consequences? Ignorance of limitations of contemporary state-of-the-art methods is hardly new [12]. Ambiguity indeed sets a meta-perspective on statistical analysis that cannot be avoided and thus needs further exploration.

Author details

Jan Peter Hessling
Kapernicus AB, Hallingsjo, Sweden

*Address all correspondence to: peter@kapernicus.com

IntechOpen

References

[1] Huxley A. Proper Studies. London: Chatto and Windus; 1927

[2] Bennett JH, Fisher RA. Statistical methods, experimental design, and scientific inference. Fisher. New York: Oxford University Press; 1995

[3] Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;**2**(8):e124. DOI: 10.1371/journal.pmed.0020124

[4] Weisberg HI. Willful Ignorance: The Mismeasure of Uncertainty. Hoboken, New Jersey: John Wiley & Sons; 2014. ISBN: 978-0470890448. ISBN: 0470890444

[5] Jaynes ET. Information theory and statistical mechanics. Physical Review. 1957;**106**(4):620

[6] Uhlmann JK. Covariance consistency methods for fault-tolerant distributed data fusion. Information Fusion. 2003; **4**(3):201-215

[7] Kalnay E. Atmospheric Modeling, Data Assimilation and Predictability. New York: Cambridge University Press; 2003. ISBN: 978-0-521-79179-3. ISBN: 978-0-521-79629-3

[8] Hessling JP. Identification of complex models. SIAM/ASA Journal on Uncertainty Quantification. 2014;**2**(1): 717-744

[9] Kay SM. Fundamentals of Statistical Signal Processing, Estimation Theory. Vol. 1. Upper Saddle River, New Jersey: Prentice Hall PTR; 1993

[10] Ljung L. System Identification—Theory for the User. Englewood Cliffs, NJ: Prentice-Hall; 1987. 519 p. ISBN 0-13-881640

[11] Feynman RP. Cargo Cult Science: Some remarks on science pseudoscience, and learning how to not fool yourself. Caltech's 1974 commencement address. Engineering and Science. 1974;10-13

[12] Fisher RA. On the mathematical foundations of theoretical studies. Philosophical Transactions of the Royal Society A. 1922;**222**:309-368. Reproduction with Author's note: https://web.archive.org/web/20051213222222; http:/www.library.adelaide.edu.au/digitised/fisher/18pt1.pdf