Remote Sens. **2013**, *5*, 1355-1388; doi:10.3390/rs5031355

OPEN ACCESS

Remote Sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Statistical Distances and Their Applications to Biophysical Parameter Estimation: Information Measures, M-Estimates, and Minimum Contrast Methods

Ganna Leonenko, Sietse O. Los * and Peter R. J. North

Department of Geography, Swansea University, Singleton Park, Swansea SA2 8PP, UK;
E-Mails: leonenkog1@cardiff.ac.uk(G.L.); p.r.j.north@swansea.ac.uk(P.R.J.N.)

⋆ Author to whom correspondence should be addressed; E-Mail: s.o.los@swansea.ac.uk;
Tel.: +44-1792-295-144; Fax: +44-1792-295-324.

Received: 16 January 2013; in revised form: 4 March 2013 / Accepted: 7 March 2013 / Published: 14 March 2013

Abstract: Radiative transfer models predicting the bidirectional reflectance factor (BRF) of leaf canopies are powerful tools that relate biophysical parameters such as leaf area index (LAI), fractional vegetation cover f_V and the fraction of photosynthetically active radiation absorbed by the green parts of the vegetation canopy (f_{APAR}) to remotely sensed reflectance data. One of the most successful approaches to biophysical parameter estimation is the inversion of detailed radiative transfer models through the construction of Look-Up Tables (LUTs). The solution of the inverse problem requires additional information on canopy structure, soil background and leaf properties, and the relationships between these parameters and the measured reflectance data are often nonlinear. The commonly used approach for optimization of a solution is based on minimization of the least squares estimate between model and observations (referred to as cost function or distance; here we will also use the terms "statistical distance" or "divergence" or "metric", which are common in the statistical literature). This paper investigates how least-squares minimization and alternative distances affect the solution to the inverse problem. The paper provides a comprehensive list of different cost functions from the statistical literature, which can be divided into three major classes: *information measures, M-estimates and minimum contrast methods*. We found that, for the conditions investigated, Least Square Estimation (LSE) is not an optimal statistical distance for the estimation of biophysical parameters. Our results indicate that other statistical distances, such as the two power measures, Hellinger, Pearson chi-squared measure, Arimoto and Koenker–Basset distances result in better estimates of biophysical parameters than LSE; in some cases the parameter estimation was improved by 15%.

Keywords: model inversion; biophysical parameter estimation; radiative transfer model; satellite data; information measures; robust statistics; minimum contrast estimation

1. Introduction

Biophysical parameters estimated from satellite data are important inputs to ecological models and land-surface models [1,2]. Various algorithms have been developed to estimate biophysical parameters from remotely-sensed reflectance data [3,4]. The forward problem, *i.e.*, to predict the reflected radiation and canopy light interactions given the structure and optical properties of the canopy and surface, is well understood and several models exist that produce realistic results [4]. The inverse problem is difficult to solve, however, because the problem is underdetermined. A commonly adopted way to overcome this is to add additional constraints or make *a priori* assumptions regarding the properties of the land surface, see [5–7].

Biophysical parameters are estimated from satellite data by inverting a sample of the of bidirectional reflectance factor, $\mathrm{BRF}(\lambda) = f(\text{Angular Geometry}, \text{Structural Parameters})$, where structural parameters (canopy properties, soil background reflectance, *etc*.) are input parameters and $\mathrm{BRF}(\lambda)$ is the model output (wavelength-dependent reflectance). Numerical solution of this inverse problem adjusts the model parameters such that model-predicted values closely match the measured values [7,8]. The match between model output and data is usually based on minimizing the sum of least squares [9].

Four approaches can be distinguished to estimate biophysical parameters from satellite data; the advantages and limitations of various approaches of biophysical parameter estimation are discussed in [10,11]. A first approach is to estimate biophysical parameters from an empirical relationship with a spectral index, see for example [1]. A second approach is to invert an analytical model; this approach puts a high demand on computing resources if the analytical model is complex. A third approach is to use machine learning, for a example by training a neural network using the inputs and outputs of a BRF model [12,13]. A fourth approach is to use LUTs. This is an attractive way to estimate biophysical parameters for various reasons. Solutions of the model can be constrained to a range of realistic input parameters, optimization is fast and the complexity of the analytical model is retained [14]. In the present study we adopt a LUT-based inversion using the FLIGHT radiative transfer model [9,15].

The estimation of biophysical parameters from satellite data is hampered by uncertainties and errors that arise from a number of sources. These include uncertainties in instrument calibration, variations in atmospheric composition or simplifying assumptions in the representation of canopy and soil background [16,17]. Errors in the representation of the canopy and soil background are of particular concern in the present study since they have non-zero mean and are not normally distributed. Outliers and nonlinearities distort the residuals and in these cases a key assumption for using LSE is violated, which is that errors have a white noise, zero mean distribution of residuals. For this reason, we investigate three broad classes of statistical distances (in the remote sensing literature referred to as cost functions, elsewhere also known as metrics, or divergence measures) that are based on different error distributions.

These classes are: *information measures of divergence* [18], *M-estimates* [19], and *minimum contrast methods* [20,21].

The first class of distance or divergence measures is referred to as *information measures*; optimization using these measures is based on minimization of distances between two probability distributions (Section 3). Thus, we need to rewrite the BRF as a probability distribution function to apply these measures. The *information measures* can be further divided into three sub-classes. The first subclass is referred to as f-divergences, a term introduced by Kullback and Leibler [22]. This class of measures is based on the distance between probability distributions (Section 3.1). The f-divergences are not bounded, *i.e.*, they range between 0 and ∞. A second subclass of divergence measures, referred to as blended measures, allows bounds to be calculated explicitly ([23] and Section 3.3). A third subclass—consisting of generalized (h, f)-divergences or superposition of two functions, see Section 3.2—is a generalization of the f-divergences [24].

The second class, of *M-estimates*, is a broad class of functions to which, among others, least squares optimization belongs. For this class of measures, the BRF is not considered a probability distribution. Within the class of *M-estimates* are a large number of functions with robust or resistant properties (Section 4).

The third class, of *minimum contrast estimates*, considers the spectral domain. We express the BRF as a spectral density function (Section 5) to apply these measures.

The present paper has two aims. The first is to provide a review of available statistical distances and divergences to date (Sections 3–5 and Appendix). Only a few of these measures have been applied to parameter estimation from remote sensing data. The second aim is to apply the distance and divergence measures to the estimation of biophysical parameters from satellite data. The availability of a large number of statistical distances gives a high degree of flexibility, since it allows model optimization for a wide range of error distributions. We illustrate this in the numerical experiments where retrieval of biophysical parameters is tested for simulated needleleaf and broadleaf forests for ground-measured BRF. The present paper tests the use of alternative distance measures on simulated observations. This allows assessment of distance measures in a well-controlled environment with known errors in estimated biophysical parameters for a wide range of simulated land-surface conditions. The method will be tested on real observations in a follow-on study [25].

The paper is organized as follows. In Section 2 the estimation of biophysical parameters from Earth Observation (EO) data is expressed in a form comparable with statistical distance theories. Sections 3–5 describe the statistical distance and divergence measures that performed best in our study. Section 6 provides a description of the BRF simulations with FLIGHT; this includes a description of the land-surface scenes and the generation of LUTs. In Section 7 the statistical distance and divergence measures are applied to the estimation of LAI, f_V and f_{APAR} by numerical inversion of the LUTs. The Appendix contains an extensive list of distance measures with references to examples of applications in the peer-reviewed literature.

We acknowledge that the range of conditions (vegetation type, simulated error distribution, land-surface properties, BRF sampling) is limited; the results of this study can therefore only be used as a guideline.

2. Statement of the Problem

The present section formulates the BRF in a way that is appropriate for the application of statistical distances and divergence measures. First we represent the following elements in the LUT: $R_i(\lambda_1, ..., \lambda_n, \bar{\theta})$ is a realization of the BRF dependent on wavelength, solar zenith angle, relative azimuth and view zenith angle, LAI, f_V, leaf area distribution, ground reflectance, *etc*. In this notation λ is the wavelength and $\lambda_1, ..., \lambda_n \in \Lambda$, $i = 1, ..., N$, where, N is the number of entries (rows) in the LUT and $\bar{\theta}$ =($\bar{\eta}$,$\bar{\zeta}$) is a vector with unknown biophysical parameters $\bar{\eta} = (\eta_1, ..., \eta_k)$ of interest to our study (e.g., LAI, f_V, f_{APAR}) and $\bar{\zeta} = (\zeta_1, ..., \zeta_r)$ is a vector with parameters that we do not need to estimate, either because they are already known (e.g., solar zenith angle, relative azimuth and view zenith angle) or because their value is obtained by other means and is not estimated in the inversion (e.g., crown shape, soil reflectance). Denoting satellite observations by $R^*(\lambda_1, ..., \lambda_n)$, we estimate the unknown parameters, the elements of the vector $\bar{\eta}^*$, by minimizing a measure that provides the best "closeness" between R in the LUT and R^*.

Let Γ be a class of measures (distances) $\Gamma(R^*(\lambda_j),\ R_i(\lambda_j, \bar{\eta}, \bar{\zeta}))$ between two BRF functions; the LUT and the observations. The classical statistical method of inversion (or estimation and finding required $\bar{\eta^*}$) of the radiative transfer model can be formulated as a semi-parametric problem

$$\bar{\eta}^* = arg \min_{\bar{\eta}} \sup_{\bar{\zeta}} \Gamma[R^*(\lambda_j, \bar{\eta}, \bar{\zeta}),\ R_i(\lambda_j, \bar{\eta}, \bar{\zeta})] \tag{1}$$

The purpose is to find the best estimate for $\bar{\eta}^*$ by solving the minimization problem (1) using different statistical distances and divergences between simulated satellite signals ("observations") and LUTs. We consider the parameters η_s^* and $\eta_{s,i}$, $1 \leq i \leq N$ closed if $|\eta_s^* - \eta_{s,i}| = min\{|\eta_s^* - \eta_{s,i}|,\ 1 \leq s \leq k\}$, $1 \leq i \leq N$. The classical approach of this minimization problem is known as LSE, which is based on the minimization of the quadratic function

$$\sum_{\lambda_j \in \Lambda} (R^*(\lambda_j) - R_i(\lambda_j, \bar{\eta}, \bar{\zeta}))^2 \rightarrow \min_{\bar{\eta}} \tag{2}$$

We consider alternative statistical distances, which can be divided into three classes. The majority of statistical distances belong to the so-called class of *information measures*. This class considers distances or measures of divergence between two probability distributions. To apply these functions to biophysical parameter estimation, the BRF must be normalized such that the sum of probabilities is 1. The expression for the LUTs becomes

$$Q = (q_1^*, ..., q_n^*) = \frac{R^*(\lambda_1)}{\sum_{\lambda_j \in \Lambda} R^*(\lambda_j)}, ..., \frac{R^*(\lambda_N)}{\sum_{\lambda_j \in \Lambda} R^*(\lambda_j)} \tag{3}$$

and for the simulated satellite observations it is

$$P_i = (p_1^i, ..., p_n^i) = \frac{R_i(\lambda_1)}{\sum_{\lambda_j \in \Lambda} R_i(\lambda_j)}, ..., \frac{R_i(\lambda_N)}{\sum_{\lambda_j \in \Lambda} R_i(\lambda_j)} \tag{4}$$

with $\sum_{l=1}^{n} q_l = 1$, and $\sum_{l=1}^{n} p_l^i = 1$ for $1 \leq i \leq N$. Thus the BRF functions can be rewritten as discrete probability mass functions by a simple normalization.

Finally, to compare the results obtained with different distance measures, the mean absolute error in parameter retrieval is defined, for example to assess the residual error in estimated LAI the following merit function is used

$$\text{Error}_{\text{LAI}} = \frac{1}{M}\sum_{i=1}^{M} |\text{LAI}_{\text{true}} - \text{LAI}_{\text{LUT}}|$$

In the next three sections we discuss the different statistical distances evaluated in the present study. The distances are applied to two reflectance distributions, P refers to the LUT reflectances and Q refers to the observed "true" reflectances. To simplify the notation we drop the index i.

3. Information Measures

Information theory was born in 1948 when Shannon [26] published his revolutionary paper motivated by the problem of efficiently transmitting information over a noisy channel. Since Mahalanobis [27] introduced the concept of distances between two probability distributions, several other distance measures have been suggested in the statistical literature and these have been referred to as measures of distance between two distributions, measures of separation, measures of discriminatory information and measures of variation-distance. While these measures were not always introduced for the same reason, they all increase when two distributions become "further away" from each other.

Divergence is an important concept in information theory and it is useful in many applications such as multimedia classification, neuroscience, optimization of the performance of density estimation methods, and cluster analysis. Distance measures also allow a wide range of tests to see if samples are from the same distribution.

Entropies are defined over the space of distributions that form the bases of independence/dependence concepts. For these reasons, Shannon's mutual information function has been increasingly utilized in the literature [28]. Shannon's relative entropy and almost all other entropies fail to be "metric", as they violate either symmetry, or the triangular rule, or both. For these reasons, it is more appropriate to refer to these entropies as measures of divergence rather than measures of distance.

Informally, entropy can be understood as "the quantity of surprise one should feel upon reading the result of a measurement". More formally, we can write: if event A occurs with probability $P(A)$, define the "information" $I(A)$ gained by knowing that A has occurred to be

$$I(A) = -\log_2 P(A)$$

The intuitive idea is that the rarer an event A, the more information we gain if we know it has occurred.

3.1. f-Divergence Information Measures

Kullback and Leibler (KL) [22] first introduced the concept of information divergences, which are non-symmetric measures between two distributions P and Q. Typically P represents the "true" distribution of the data and Q represents a model or an approximation of P. In information theory, KL divergence can be interpreted as cross Shannon entropy. This class has been extended in many directions since its initial application in decoding schemes and in signal processing. In particular, Rényi proposed

a generalization of Shannon entropy [29], one of a family of functionals for quantifying the diversity, uncertainty or randomness of a system. The Rényi entropies are important in ecology and statistics as indices of diversity. Later, KL and Rényi related divergences were included in a broader class of divergences called f-divergences, introduced by Csiszár [30]. This class can be formulated as follows.

A general class of divergence measures is given by

$$\Gamma_F[P,Q] = \sum_{l=1}^{n} F(p_l, q_l) \tag{5}$$

where

- $0 \leq p_1, ...p_n \leq 1, 0 \leq q_1, ...q_n \leq 1, \sum_{l=1}^{n} q_l = 1$ and $\sum_{l=1}^{n} p_l = 1$;
- $F(p,q)$ is a strictly convex function of p so that $\Gamma_F(P,Q)$ is a strictly convex function of $0 \leq p_1, ..., p_n \leq 1$;
- For fixed Q, $\Gamma_f(P,Q)$ attains its unconstrained global minimum when $p_l = q_l$ for all l, *i.e.*, if $P = Q$;
- for a given strictly convex twice differentiable function $f(.)$ we define

$$\Gamma_f[P,Q] = \sum_{l=1}^{n} q_l f\left(\frac{p_l}{q_l}\right) \tag{6}$$

Note that the global minimum value is equal to $f(1)$. Thus for a given f

$$D_f[P,Q] = \Gamma_f[P,Q] - f(1) \geq 0 \; and \; = 0 \iff P = Q$$

Many measures were added to this class from different areas of science and new divergences are still being discovered. Here we present some of these f-divergence measure, see [31] and additional list can be found in Section A.1 of the Appendix.

1. Let $f(x) = x\ln(x)$ and $f(1) = 0$. The corresponding measure is

$$D_f[P,Q] = \sum_{l=1}^{n} p_l \ln\left(\frac{p_l}{q_l}\right) \tag{7}$$

   This measure is called the KL divergence; it also corresponds to the maximum likelihood distance. It has wide application in code theory, signal processing, data compression, data storage and data communication as well as others areas of science.

2. Let $f(x) = \frac{1}{x}$ and $f(1) = 1$ then

$$D_f[P,Q] = \sum_{l=1}^{n} \frac{q_l^2}{p_l} = \sum_{l=1}^{n} \frac{(q_l - p_l)^2}{p_l} \tag{8}$$

   This measure is called the Pearson chi-square. This measure is used in the chi-squared test first proposed by K. Pearson.

3. χ^α—Vajda divergence corresponds to the function $f(x) = |x-1|^\alpha$ and $f(1) = 0$, where $\alpha \geqslant 1$, then

$$D_f[P,Q] = \sum_{l=1}^{n} |p_l - q_l|^\alpha q^{1-\alpha} \tag{9}$$

4. Let $f(x) = (\sqrt{x} - 1)^2$ and $f(1) = 0$, then

$$D_f[P, Q] = \sum_{l=1}^{n} q_l \left(\sqrt{\frac{p_l}{q_l}} - 1 \right)^2 = \sum_{l=1}^{n} (\sqrt{p_l} - \sqrt{q_l})^2 \tag{10}$$

   is called the squared Hellinger measure. It is bounded full distance and has been applied to parameter estimation in censored models (where the variable of interest is only observable under certain conditions) as well as many other areas of science.

5. It can be generalized in the following form to give more flexibility on parameter estimation. Assume $f(x) = (x^\alpha - 1)^{1/\alpha}$ and $f(1) = 0$, then

$$D_f[P, Q] = \sum_{l=1}^{n} (p_l^{1/2j} - q_l^{1/2j})^{2j}, \; j = 1, 2, 3... \tag{11}$$

6. Let $f(x) = (1 - x)^{2j}$, $f(1) = 0$, then power divergence has the form

$$D_f[P, Q] = \sum_{l=1}^{n} q_l \left(1 - \frac{p_l}{q_l} \right)^{2j}, \; j = 1, 2, 3... \tag{12}$$

7. Power divergence measures [32] with minimum at zero. This class was introduced to unite efficiency with robust properties; the class is also Fisher consistent

$$D_f[P, Q] = \sum_{l=1}^{n} p_l \frac{\{[p_l/q_l]^\alpha - 1\}}{\alpha(\alpha + 1)} \tag{13}$$

   which gives for $\alpha = -2, -1, -1/2, 0, 1$ the following already known measures: the Neyman chi-squared measure divided by 2, the Kullback–Leibler divergence, the twice-squared Hellinger distance, the likelihood disparity, and the Pearson's chi-squared divided by 2.

3.2. Other Divergence Measures between Two Probability Distributions.

There are another two important sub-classes based on the divergence between probability distributions that do not belong to the class of f-divergences. These sub-classes are referred to as (h, f)-measures and f-entropy measures.

The first of these sub-classes, (h, f)-measures, was introduced by [33]. It can be written in the following form $D_f^h[P, Q] = h(D_f[P, Q])$, where h is a differentiable increasing function mapping from $\left[0, f(0) + lim_{t\to\infty} \frac{f(t)}{t}\right] \to [0, \infty]$. Under different assumptions, it is shown that the asymptotic distributions of the (h, f)-divergence statistics are either normal or chi-square. These divergences were developed for hypothesis testing on multinomial populations and to test goodness of fit and independence. This class is based on the superposition of two functions and it gives a large degree of flexibility to deal with outliers.

Here is one of the examples of these measures. Additional list can be found in Section A.2 of the Appendix.

Rényi divergence with

$$h(x) = \frac{1}{\alpha(\alpha - 1)} log(\alpha(\alpha - 1)x + 1); \; f(x) = \frac{x^\alpha - \alpha(x - 1) - 1}{\alpha(\alpha - 1)}; \; \alpha \neq 0, 1$$

$$D_f^h[P,Q] = \frac{1}{\alpha(\alpha-1)} log\left(\left(\sum_{l=1}^{n} q_l(p_l/q_l)^\alpha - \alpha(p_l - q_l) - q_l\right) + 1\right) \tag{14}$$

The second subclass is referred to as entropy measures and can be introduced as follows. Let X be a random variable with probability distribution P. Shannon's entropy [26] has the form

$$H(X) = H(P) = -\sum_{l=1}^{n} p_l \log p_l \tag{15}$$

while the cross-entropy is

$$H(P/Q) = -\sum_{l=1}^{n} p_l \log q_l \tag{16}$$

It is easy to verify that the KL divergence Equation (7) is related to Shannon's entropy, *i.e.*, $D_f[P,Q] = H(P) - H(P/Q)$.

In order to present a systematic way of studying the different entropy measures, Burbea and Rao introduced the so-called f-entropies, by

$$H_f(X) = H_f(P) = \sum_{l=1}^{n} f(p_l) \tag{17}$$

where $f : (0,\infty) \to R$ is a continuous concave function and $f(0) = \lim_{t\downarrow 0} f(t) \in (-\infty, \infty)$. It turned out that some important entropy measures cannot be written as f-entropy. For this reason [24], defined the (h, f)-entropy as follows,

$$H_f^h(X) = H_f^h(P) = h(H(P)) \tag{18}$$

where either $f : (0,\infty) \to R$ is concave and $h : R \to R$ is differentiable and increasing, or $f : (0,\infty) \to R$ is convex and $h : R \to R$ is differentiable and decreasing.

Based on the concavity property of the (h, f)-entropy, new generalization was introduced in [34]:

$$D_f^h[P,Q] = H_f^h\left(\frac{P+Q}{2}\right) - \frac{H_f^h(P) + H_f^h(Q)}{2} \tag{19}$$

These measures of divergence have been introduced to present systematic ways to study different entropy measures. They are used in applications that are associated with random variables with finite support in genetic diversity between populations, the study of taxonomy in biology and to test if populations are homogeneous in genetics and for the analysis of discriminant techniques.

An example of these measures can be seen below and additional list can be found in Section A.2 of the Appendix.

Arimoto (1971)

$$f(x) = x^{1/\alpha},\ h(x) = \frac{1}{\alpha - 1}(x^\alpha - 1),\ \alpha > 0,\ \alpha \neq 1$$

$$D_f^h(P,Q) = \left(\frac{1}{\alpha-1}\right)\left[\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{1/\alpha}\right)^\alpha - \frac{1}{2}\left[\left(\sum({p_l}^{1/\alpha})\right)^\alpha + \left(\sum({q_l}^{1/\alpha})\right)^\alpha\right]\right] \tag{20}$$

3.3. Blended f-Disparities

A third group of divergences is referred to as blended divergences. Lindsay [32] found that inference based on statistics of type f-divergence (obtained by replacing either one or both probability distributions by suitable estimators) requires either bounded differentiability of f or boundedness of f itself. He introduced a new class of divergences by the modification of weights inside the integral expression of Pearson's chi-squared divergence called "blended weight chi-squared disparity"—BWCS(β) and "blended weight Hellinger disparity"—BWHD(β), $\beta \in [0,1]$. In general all these new classes of disparities have the following common property. If the blending parameter is equal to the limiting values $\beta = 0$ or $\beta = 1$, then the two original divergences on which the blend was based are achieved in the class of blended divergences. Definitions and theorems can be found in [23].

All blended f-disparities have been used in goodness-of-fit tests in medical statistics and are shown to be an excellent compromise between the Pearson's chi-square and the log likelihood ratio tests.

To illustrate the theory of blended divergences, we give several examples below.

Blended weighting scheme that generalizes Hellinger distance:

$$D_\alpha(P,Q) = \frac{1}{2}\sum_{l=1}^{n} \frac{(p_l - q_l)^2}{(\alpha\sqrt{p_l} + (1-\alpha)\sqrt{q_l})^2},\ \alpha \in (0,1) \tag{21}$$

4. Nonlinear Regression and M-Estimates

Robust regression is a form of regression analysis designed to circumvent some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as LSE, have favorable properties if their underlying assumptions are true, but can give misleading results if those assumptions are violated. In cases where errors are not normally distributed, outliers occur, or ordinary LSE assumptions are violated in some other way, the validity of the regression results is compromised if a non-robust regression technique is used. *M-estimators* (see for example [35]) form a broad class of estimators that exhibit certain robust properties. Estimates with robust regression methods can be more stable with respect to anomalous errors.

We use *M-estimates* to the estimation of biophysical parameters from reflectances as follows. Let us consider the BRF $R(\lambda_j, \bar{\theta})$ as a nonlinear regression function $g(\lambda_j, \bar{\theta})$ of its parameter set, which is observed at wavelength $\lambda_j \in \Lambda$ with some noise ϵ_j of complicated nature. The observations set $R^*(\lambda_j) = x_j$ can be represented as follows

$$x_j = g(\lambda_j, \bar{\theta}) + \epsilon_j,\ j = 1, ..., n \tag{22}$$

where $E\epsilon_j = 0$, *i.e.*, the expectation is that ϵ_j is random noise with zero mean, not Gaussian in general.

In our case we are interested in the unknown parameter vector of interest $\bar{\eta}$ =(LAI, f_V, f_{APAR}). We use the *M-estimates* generated by a function of loss $\rho(x)$, $x \in \mathbb{R}$, see [19,35], for some motivation details and properties of these estimates.

The *M-estimate* of the unknown parameter vector $\bar{\eta}$ obtained from the observations x_j $j = 1, ..., n$ is given by the solution of the minimization problem:

$$Q_n(\bar{\eta}) = \sum_{j=1}^{n} \rho(x_j - g(\lambda_j, \bar{\eta})) \rightarrow \min_{\bar{\eta}} \tag{23}$$

The function $\psi(x) = \rho'(x)$ is called the score function and the minimization problem (Equation (23)) can be written in the following equivalent way:

$$\sum_{j=1}^{n} \psi(x_j - g(\lambda_j, \bar{\eta}))\nabla g_{\bar{\eta}}(\lambda_j \bar{\eta}) = 0. \tag{24}$$

The classical LSE corresponds to the case $\rho(x) = x^2, \psi(x) = 2x$. In this case random noise ϵ_j are i.i.d.r.v. that have a Gaussian distribution. In general for nonlinear regression, ϵ_j may be independent but not necessarily Gaussian or even non-Gaussian dependent random variables. It is well known that LSE regression methods are consistent, asymptotically normal and asymptotically efficient. However, when the density function of errors is non-Gaussian or even has a non-symmetric skewed distribution, LSE estimates are no longer efficient and their application can result in large losses of efficiency. Robust methods replace the sum of squares by more suitable loss functions.

The following examples belong to the class of *M-estimates* and can be found in [35–37] and others. The full list of M-estimates can be found in Section A.2 of the Appendix.

(1) If errors are normally distributed $f(x) = (1/\sqrt{2\pi})exp(-x^2/2)$, then

$$\rho(x) = x^2, \psi(x) = x \tag{25}$$

is LSE and it is non-robust. It is used widely in many application of remote sensing, biology, economy and other areas of science since it has nice properties described above.

(2) The function

$$\rho_c(x) = \begin{cases} cx, & x \geq 0 \\ (c-1)x, & x < 0 \end{cases} \tag{26}$$

defines the class of Koenker–Basset estimators [38], $(0 < c < 1)$. In contrast to classical methods based on least-squares residuals, this measure is robust and has an explicit probabilistic meaning. It was introduced for the estimation of an unknown parameter $\bar{\eta}$ in the nonlinear regression model (Equation (23)) when the samples are independent, but errors $\epsilon_j = x_j - g(\lambda_j, \bar{\eta})$ have some skewness (non-symmetric) property. The empirical quantile function may be defined in terms of solutions to a simple optimization problem. Explicitly,

$$\hat{Q}_\epsilon(c) = inf\left\{ y | \sum_{j=1}^{n} \rho_c(\epsilon_i - y) = min \right\} \tag{27}$$

where ρ_c-function (Equation (26)). One can also interpret this as $(E\epsilon_j = 0)$

$$\begin{cases} P\{\epsilon_i \geq 0\} = c \\ P\{\epsilon_i < 0\} = 1 - c \end{cases} \tag{28}$$

This distance measure has been applied in econometrics to study wage distributions and to study linear regression quantiles on censored data.

5. Minimum Contrast Estimation

To apply *minimum contrast measures* to our problem, we interpret the BRF as a spectrum or spectral density of some stochastic process. The basic idea behind a *minimum contrast estimator* is to minimize the distance (contrast) between a parametric model and a non-parametric spectral density. The estimates obtained are first-order efficient and are also attractive because they have robust properties. This class of estimates is close to the class of quasi-likelihood estimators, where instead of independence (which does not hold for many cases) we use asymptotical independence, as discussed below.

We adopt the following terminology from time series analysis to interpret our observations as measurements in the spectral domain [39]. Let $\{Z_t\}$ be a stationary processes with spectral density $f(\lambda) = f_\theta(\lambda)$, $\lambda \in \Lambda \subseteq R$, with expectation m, and $\theta \in \Theta \subset \mathbb{R}^p$. Our aim is to estimate the unknown parameter θ, that is, to identify the true value of spectrum $f_{\theta^*}(\lambda)$.

To implement this idea we consider the so-called quasi-likelihood method [40]. The BRF observations in the spectral domain are written in a form, $I_n(\lambda_j)$, $\lambda_j \in \Lambda = \{\lambda_1, ..., \lambda_n\}$, where

$$I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n}(Z_j - m)e^{ij\lambda}\right|^2, \ \lambda \in \Lambda \tag{29}$$

This is the so-called periodogram non-parametric estimation of spectral density $f(\lambda)$.

Under some general conditions [39], at the Fourier frequencies $\lambda_j \in \Lambda$, the random variables $I_n(\lambda_j)$ are asymptotically independent and have an exponential distribution, that is

$$\lim_{n\to\infty} P\{I_n(\lambda_j) \le u\} = 1 - e^{-\frac{u}{f(\lambda_j,\theta)}}, \ u \ge 0 \tag{30}$$

The pdf that corresponds to the distribution function in the right hand side of Equation (31) takes the form: $\frac{1}{f(\lambda_j,\theta)}e^{-\frac{u}{f(\lambda_j,\theta)}}$.

Thus, one can construct a quasi-likelihood function or its logarithm

$$logL(I_n(\lambda_1), ...I_n(\lambda_n)) = log\prod_{j=1}^{n}\frac{1}{f(\lambda_j,\theta)}e^{-\frac{I_n(\lambda_j)}{f(\lambda_j,\theta)}}$$

$$= -\sum_j\left[logf(\lambda_j,\theta)) + \frac{I_n(\lambda_j)}{f(\lambda_j,\theta)}\right] \tag{31}$$

which has to be maximized in order to estimate θ. It means that we need to minimize the so-called Whittle functional

$$Q_n(\theta) = \sum_j\left[logf(\lambda_j,\theta)) + \frac{I_n(\lambda_j)}{f(\lambda_j,\theta)}\right] \tag{32}$$

and the quasi-likelihood estimate is

$$\hat{\theta}_n = \underset{\theta\in\Theta}{Argmin} Q_n(\theta)$$

The Whittle estimator was also extended to cover correlated signal-plus-noise models, providing a formal asymptotic distribution theory specifically tailored for parameter estimation. This approach was first applied in time series for exponential volatility models; it then caught attention in financial econometrics and in related fields. These models are able to represent some of the stylized features of

financial returns, such as uncorrelation in levels but strong dependence in squares and log-squares and leverage effect.

The Whittle estimator belongs to a class of more general estimates known as minimum contrast estimates, see [20,21]. To demonstrate the idea, let us assume that the true value of unknown parameter $\theta^* \in int\Theta$, the interior of Θ. A contrast function for θ^* is a deterministic function $F(\theta^*, \theta)$, $F_{\theta^*} : \Theta \to \mathbb{R}_+$, which has a unique minimum, $\theta = \theta^*$.

A contrast process for F_{θ^*} is a sequence of random variables $U_n(\theta)$, $n = 1, 2...$ such that the ergodic like condition holds for some $U(\theta)$ in probability:

$$\lim_{n\to\infty} [U(\theta) - U_n(\theta^*)] = F(\theta, \theta^*)$$

The minimum contrast estimator is a value of θ for which the function $U_n(\theta)$ takes its minimum, or

$$\hat{\theta}_N = Argmin_{\theta\in\Theta} U_n(\theta)$$

Under some sets of conditions the minimum contrast estimators are consistent, see [41]. Often the contrast function can be chosen as a distance $L(f_\theta, g)$ between two spectral densities $f_\theta(\lambda)$ and $g(\lambda)$, which can be written in the form:

$$L(f_\theta, g) = \int_\Lambda K\{f_\theta(\lambda)/g(\lambda)d\lambda \tag{33}$$

where $K(x)$ is a three times continuously differentiable function on $(0, \infty)$ and has a unique minimum at $x = 1$. The contrast process in practice can be approximated as follows:

$$U_n(\theta) = \int_\Lambda K\{f_\theta(\lambda)/I_n(\lambda)d\lambda \approx \sum_{\lambda_j\in\Lambda} K\left(\frac{f_\theta(\lambda_j)}{I_n(\lambda_j)}\right) \tag{34}$$

The following examples of distances $L(f_\theta, g)$ are widely used in parametric estimation in time-series analysis in frequency domain, in particular for autoregressive and moving average models.

One of the example can be seen below and the full list of such distances can be found in Section A.3 of the Appendix.

Let $K(x) = logx + \frac{1}{x}$. This criterion is equivalent to the quasi-Gaussian maximum likelihood (73) and has the following form

$$L(f_\theta, g) = \sum_{\lambda_j\in\Lambda} \{log(f_\theta(\lambda_j)/g(\lambda_j)) + g(\lambda_j)/f_\theta(\lambda_j)\} \tag{35}$$

Note that function $K(x)$ have a unique minimum at $x = 1$. To find a minimum at zero we need to subtract $K(1)$ under the sum from each of the functions. In practical applications we use the notation from Section 2, *i.e.*, in the above methodology the abstract parameter θ is replaced by the vector parameter of our interest $\bar{\eta} = (\eta_1, ..., \eta_k)$.

Additional cost functions from the literature [42–48] are summarized in the Appendix.

6. Methodology

6.1. Radiative Transfer Modeling

The performance of the distance measures in terms of retrieving biophysical parameters (LAI, f_{APAR}, f_V) from reflectance values was tested on simulations. This approach has the advantage that the estimated biophysical parameters can be compared directly with model input parameters and that therefore errors associated with the use of different distance measures can be established with accuracy. Moreover, we can test distance measures on a large number of simulations and this provides a good indication of their robustness.

Simulations were carried out similar to the approach taken by Prieto-Blanco *et al.* [9]. We used two models in conjunction to simulate a set of the ground BRF observations and to generate the LUTs. Simulations were carried out for 3 needleleaf and 2 broadleaf scenes. We used PROSPECT [49] to simulate light scattering and absorption by leaves, and FLIGHT [15] to simulate light scattering and absorption by vegetation canopies. The models used are state of the art and provide realistic simulations of the interaction of solar radiation with the vegetation canopy and the soil.

PROSPECT [49] calculates leaf transmittance and reflectance from 400 to 2,500 nm. In PROSPECT4 each leaf is considered as a stack of N absorbing plates with rough surfaces giving rise to scattering of light. Absorption is calculated as the linear summation of the concentrations of chlorophyll, water, and dry matter, each with their specific absorption coefficients [49]. The PROSPECT input parameters are described in Table 1. The inputs include: N, the leaf structure parameter; C_{ab}, the chlorophyll $a+b$ concentration ($\mu g/cm^2$); Cw, the equivalent water thickness (g/cm^2); Cm, the dry matter content (g/cm^2). Chlorophyll content (C_{ab}) in leaves is linked to the maximum photosynthetic capacity of vegetation and varies with leaf development stage, productivity, stress and nitrogen levels. For the LUTs of conifers, a maximum and minimum value for C_{ab} was entered to reflect a range of conditions.

Table 1. PROSPECT input parameters. N is the leaf structure parameter; C_{ab} are the chlorophyll $a+b$ concentrations ($\mu g/cm^2$); Cw is the equivalent water thickness (g/cm^2); and Cm is the dry matter content (g/cm^2).

	CONIFER			**BROADLEAF**	
	OBS	**OJP**	**YJP**	**Beech**	**Oak**
N	2.47	2.55	2.55	1.43 1.5 1.6	1.61 1.97 2.64
Cab	29 19.39 27.56	27.07 13.10 24.27	21.89 19 29.03	44.7	65.1
Cw	0.04	0.01	0.03	0.02	0.008
Cm	0.028	0.012	0.012	0.003	0.006

FLIGHT [15] is a 3D radiative transfer model for light interaction with vegetation canopies using Monte Carlo simulation of photon transport. The original model traced the photons' trajectories forwards from the source until they were absorbed in the canopy or left the canopy boundary. Subsequent improvements include calculation of paths back from any view direction to the intercepted surface

facets [50,51], simulation of fine angular resolution, simulation of photosynthesis and simulation of LiDAR signals [52]. A hybrid representation is used to model the discontinuous nature of the forest canopy. Large-scale structure is represented by geometric primitives defining shapes and positions of the tree crowns and trunks, here estimated from a statistical distribution. Within each crown, foliage is approximated by structural parameters of area density, angular distribution and size and optical properties of reflectance and transmittance. These parameters are approximated as homogeneous within each boundary, but may vary between crowns. Simulation of 3D photon trajectories allows accurate evaluation of multiple scattering within crowns, and between distinct crowns, trunks and ground surface. FLIGHT simulations have previously been compared with other 3D canopy radiative transfer models as part of the Radiation Model Intercomparison (RAMI) project [4]. The recent analysis within RAMI of six selected 3D models, including FLIGHT, showed dispersion within 1% over a large range of canopy descriptions, see [53].

Radiation was simulated in 15 spectral bands (500, 560, 630, 690, 700, 740, 790, 830, 870, 1,035, 1,200, 1,250, 1,650, 2,100, 2,250 nm). A previous study [54] suggested such a selection of bands could provide approximately 90% of the information about the land surface that is provided by a full spectrum, although this study was based on field spectroscopy. The set is chosen here to demonstrate the retrieval method and selection of error metrics, but the method is applicable to any set of bands or potential view directions, where the study should be repeated to determine optimal error metrics.

6.2. Sites

The simulations were carried out on two main types of forest: conifer and broadleaf forests. Three conifer representatives were chosen from the former BOREAS sites [55], each characterized by a different dominant species: the Old Black Spruce (*Picea Mariana*) site (OBS), the Old Jack Pine (*Pinus Banksiana*) site (OJP) and the Young Jack Pine (*Pinus Banksiana*) site. Vegetation of these sites has a complex structure, needles show a high degree of clumping and there is mutual shadowing by crowns. These sites are therefore known to pose a challenge to biophysical parameter estimation. Detailed crown, leaf and soil background measurements are available for these sites as these have been extensively studied in [56,57]. Chlorophyll content in coniferous canopies has been estimated in [58]. Changes in leaf chlorophyll produce large differences in leaf reflectance and transmittance spectra, therefore three values of C_{ab} were used to obtain a wide range of possible values [59], (Table 1). Broadleaf simulations were carried out for an oak and beech forest since these are among the most important species for the European forestry [60,61].

Table 1 shows the leaf optical properties for the conifer and broadleaf forests that were used in PROSPECT; values are based on [62]. Tables 2 and 3 show the vegetation structure parameters and the angular configurations for the FLIGHT model. Five characteristic soil spectra from the Purdue spectral library were selected (Table 2) [63,64].

Table 2. FLIGHT input parameters. The column "range" represents the minimum and maximum values of the input parameters; column "step" is the increment for the LUT; column "Observed" represents the range over which a random number "r.n." is selected to generate satellite observed BRF, $R^*(\lambda_1, ..., \lambda_n)$. Thus the simulations and LUT are generated from different input parameters; in particular, different values are used for viewing geometry, soil reflectance, fractional cover, and leaf area index (Section 6.4).

Parameter	**Range**	**Step**	**"Observed"**
Solar zenith angle	30°–70°	10°	r.n.∈30°–70°
View zenith angle	0°–60°	10°	r.n ∈0°–60°
Relative azimuth angle	0°–180°	30°	r.n.∈0°–180°
Fraction of green leaves	0.8	-	same
Fraction of shoot material	0.05	-	same
Fraction of bark in foliage	0.15	-	same
Leaf angle distribution	Spruce leaf, Spherical	-	same
Soil roughness index	0	-	0
Soil Reflectance	sandy loam		drummer2, jal, lonrina, onaway, talbott
Frac. cover by trees	0.1–0.9	0.1	r.n.∈0.0–0.9
LAI	0–7, $LAI \leq 8FC$	1	r.n.∈0.0–7.0

Table 3. FLIGHT input parameters: Crown shapes, where "c" represents cone shape and "e" represents ellipsoid shape.

Parameter	**Conifer forest**			**Broadleaf forest**	
Crown shape	cone			ellipsoid	
type of forest	OBS	OJP	YJP	Beech	Oak
crown shape	"c"	"c"	"c"	"e"	"e"
Crown radius (m)	0.45	1.3	0.85	1.2	2.6
Crown center to top dist (m)	9	7.2	4	4.2	3.2
Minimum height to first branch (m)	0.49	6.9	0.49	6.4	7.1
Maximum height to first branch (m)	0.51	7.1	0.51	10.2	9.2

6.3. Generation of Look-Up Tables

The LUT contains a total of $N = 90,404$ entries of BRF reflectances. These are reflectance values calculated for parameters obtained at regular intervals of solar zenith angle, view zenith angle, relative azimuth angle, LAI, f_V and C_{ab}, see Table 1. Crown shape parameters can be found in Table 3 and

f_{APAR} is obtained by summing individual f_{APAR} values at each band times weighted by the fraction of downwelling light within the band:

$$f_{\text{APAR}} = \sum_{i=1}^{n} F_{a,i} W_i$$

where $F_{a,i}$ is the mean fraction of radiant energy absorbed by canopy at band i, calculated from FLIGHT output, and W_i are the weights, see also [9].

6.4. Simulation of Observations

We simulated BRF values using the coefficients in Tables 1–3. For each realization the observation tables contain a total of M = 5,000 entries for the three conifer sites (OBS, OJP, YJP) and M = 5,000 for the broadleaf sites (beech and oak). The tables were simulated similar to the calculation of the LUT but using a random selection from the range of parameters in Table 2, last column (LAI, f_V, angular geometry, atmospheric aerosol depth). Note that the parameters from Tables 1 and 3 stay the same. The observations are simulated from a wider range of conditions that are not present in the conifer and broadleaf LUTs (soils, leaf area distribution). Further errors due to sensor noise and calibration are not considered here, but have been examined in previous studies [9,65] and should be considered prior to application to particular instruments. The tables of "observed" ground reflectances were constructed to contain complex errors for needleleaf and broadleaf forests. These errors, originating from the mismatch between LUT and "observations", represent conditions encountered in real applications where we have to match a model representing only a selection of conditions with richly varied real world conditions.

Canopy reflectance models demonstrate increasing sensitivity to soil reflectance at lower vegetation cover. Soil reflectance is one of the most sensitive parameters in canopy reflectance models [17]. Errors are biased and unsymmetrical and for this reason we expect some robust distances to perform better than LSE. This effect is more pronounced at lower values of LAI (<3).

Our simulations are necessarily restricted and represent only a subset of all parameters that can be varied within Prospect and FLIGHT. We believe that these simulations of errors are more realistic than the commonly adopted method of adding Gaussian noise to the spectrum. Errors due to incorrect assumptions on soil or leaf spectral properties will be spectrally correlated.

7. Results

The statistical distances listed in Sections 3–5 were evaluated as follows. Reflectances in the LUTs were matched to "observed" reflectances. For each case, we matched *one* entry in the "observations" table consisting of M = 5,000 spectral bands with *one* entry in the LUT for each type of forest. The inversion finds parameters for the nearest angular geometry in the LUT, *i.e.*, the angular geometry is known. Other parameters are assumed to be unknown. The performance of the statistical distances was then assessed on the biophysical parameter estimated (LAI, f_V,f_{APAR}). Some statistical distances allow parameters that govern the shape of the error distribution to be varied, thus the choice of these parameters leads to another optimization problem in itself. For these cases we tested a range of parameters and chose the parameter that minimizes the error in estimated biophysical parameters. All distances were tested by comparing the estimated biophysical parameters with the *a priori* known parameters.

For the purpose of clarity we present a selection of results that consists of the best performing measures from each of the three classes of statistical distance measures. The results show significant variability in retrieval accuracy depending on the chosen divergence measure. Overall, the optimal measures in each case show an improvement over using LSE, see Tables 4 and 5.

Table 4. Summary statistics for the performance of different distance measures in estimating biophysical parameters from ground reflectance data (BRF); reflectances were simulated for broadleaf trees.

Distance	**Type**	**Parameters**	**Error** LAI	**Error** f_V	**Error** f_{APAR}
Equation (7) Kullback–Leibler	inf. meas.	–	0.63	0.25	0.10
Equation (10) Hellinger	inf. meas.		***0.61***	0.26	***0.09***
Equation (12) power	inf. meas.	$j = 4$	0.69	***0.24***	0.12
Equation (11) Gen Hellinger	inf. meas.	$j = 2$	0.66	0.26	0.10
Equation (14) Rényi	inf. meas.	$\alpha = 0.5$	0.63	0.26	0.10
Equation (21) Blended Gen Helling	inf. meas.	$\beta = 0.9$	0.63	0.26	0.10
Equation (20) Arimoto	inf. meas.	$\alpha = 0.8$	0.62	0.26	0.10
Equation (25) LSE	M-estim	–	**1.79**	**0.34**	**0.21**
Equation (26) Koenker -B.	M-estim	$\alpha = 0.99$	0.92	0.29	0.12
Equation (35)	min. cont. meth.	–	1.01	0.26	0.16

Table 5. Summary statistics for the estimation of biophysical parameters from ground reflectance values (BRF) for needleleaf canopies.

Distance	**Type**	**Parameters**	**Error** LAI	**Error** f_V	**Error** f_{APAR}
Equation (9) Vajda	inf. meas.	$\alpha = 3$	0.74	0.21	0.08
Equation (8) Pearson χ^2	inf. meas.	–	0.69	0.22	***0.07***
Equation (13) power	inf. meas.	$\alpha = -5$	***0.68***	0.22	***0.07***
Equation (12) power	inf. meas.	$j = 4$	0.77	***0.19***	0.08
Equation (14) Rényi	inf. meas.	$\alpha = 0.5$	0.70	0.22	0.078
Equation (21) Blend Helling	inf. meas.	$\beta = 0.9$	0.69	0.22	0.076
Equation (25) LSE	M-estim	–	**1.35**	**0.30**	**0.16**
Equation (26) Koenker -B.	M-estim	$\alpha = 0.2$	1.29	0.29	0.16
Equation (35)	min. cont. meth.	–	0.95	0.28	0.12

Table 4 shows the best distance measures (bold italic) for estimating biophysical parameters on broadleaf forests from BRF as well as results obtained with LSE (bold text). Table 5 shows the same but for estimating biophysical parameters on conifer forest.

The improvement obtained in estimating biophysical parameters for broadleaf forest is further illustrated in Figure 1. In Figure 1 residual errors in broadleaf biophysical parameters obtained from BRF reflectances by using LSE are compared with errors obtained by Hellinger (Equation (10)) and power divergence (Equation (12)). The range of errors (maximum and minimum errors) for these two methods

is the same, however the alternative divergence shows a marked improvement in the error distribution for all biophysical parameters.

Figure 1. Comparison of residual errors resulting from the application of two different distance measures used to estimate biophysical parameters for broad leaf canopies with a look-up table (LUT). Biophysical parameters, including leaf area index, LAI, fraction of photosynthetically active radiation absorbed by the canopy, f_{APAR}, and vegetation cover fraction, f_V, are estimated from simulated ground reflectance values. The rows from top to bottom show the comparison of residual errors in estimating LAI with the least squares estimate (LSE) and Hellinger Equation (10) (top row), of residual errors in f_V with LSE and the power divergence Equation (12) with power 1 and $j = 4$ (centre row) and of residual errors in f_{APAR} with LSE and the Hellinger equation (bottom row). The columns from left to right show frequency distributions of residual errors using LSE as a cost function (left column) and the residual errors using either a Hellinger Equation (10) for LAI and f_{APAR} or a power divergence Equation (12) with power 1 and $j = 4$ for f_V (centre column). The right column shows the quantile–quantile plots comparing the frequency distributions of the left and centre columns expressed as absolute errors. The errors in biophysical parameters associated with the alternative cost functions are smaller and better behaved (more symmetrical and smaller bias).

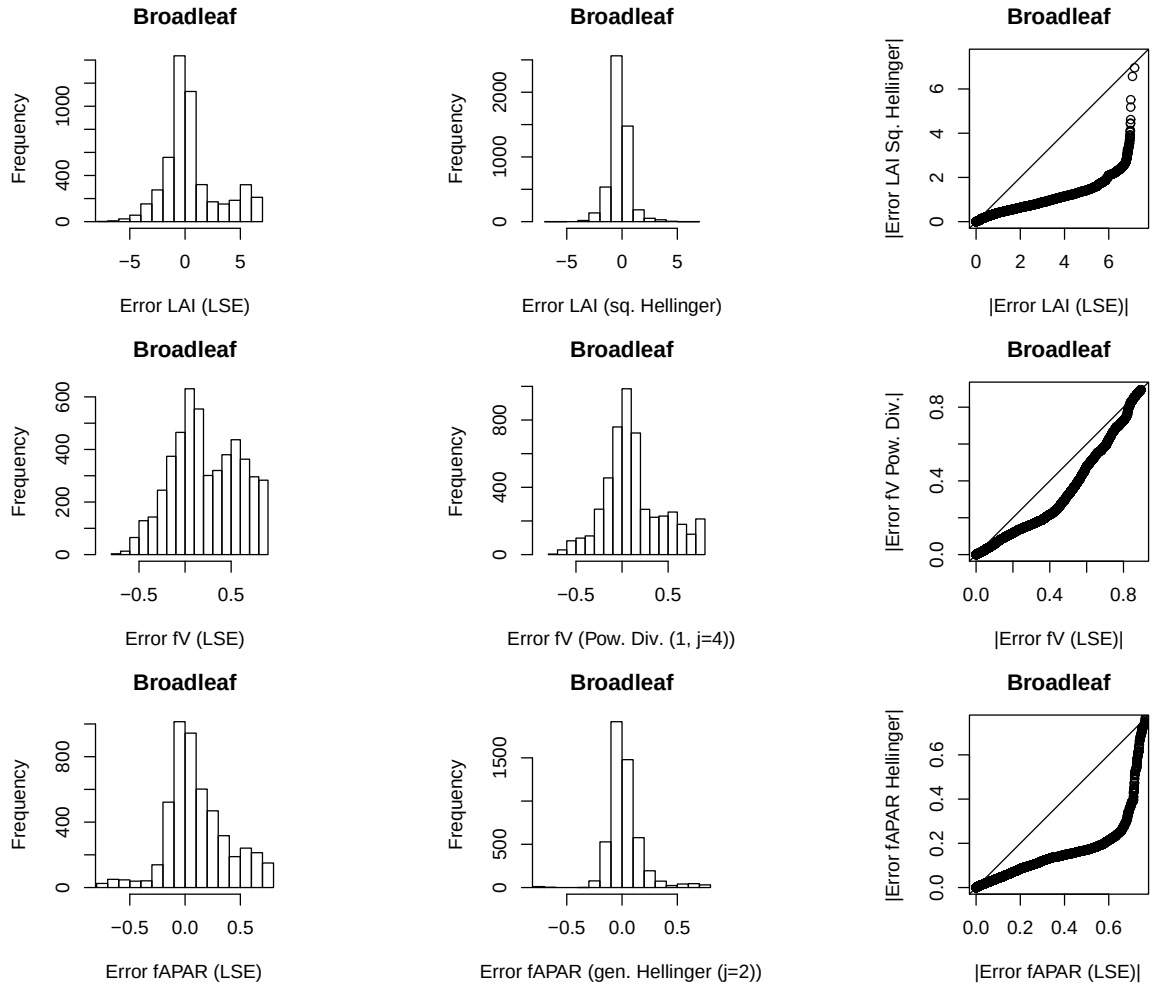
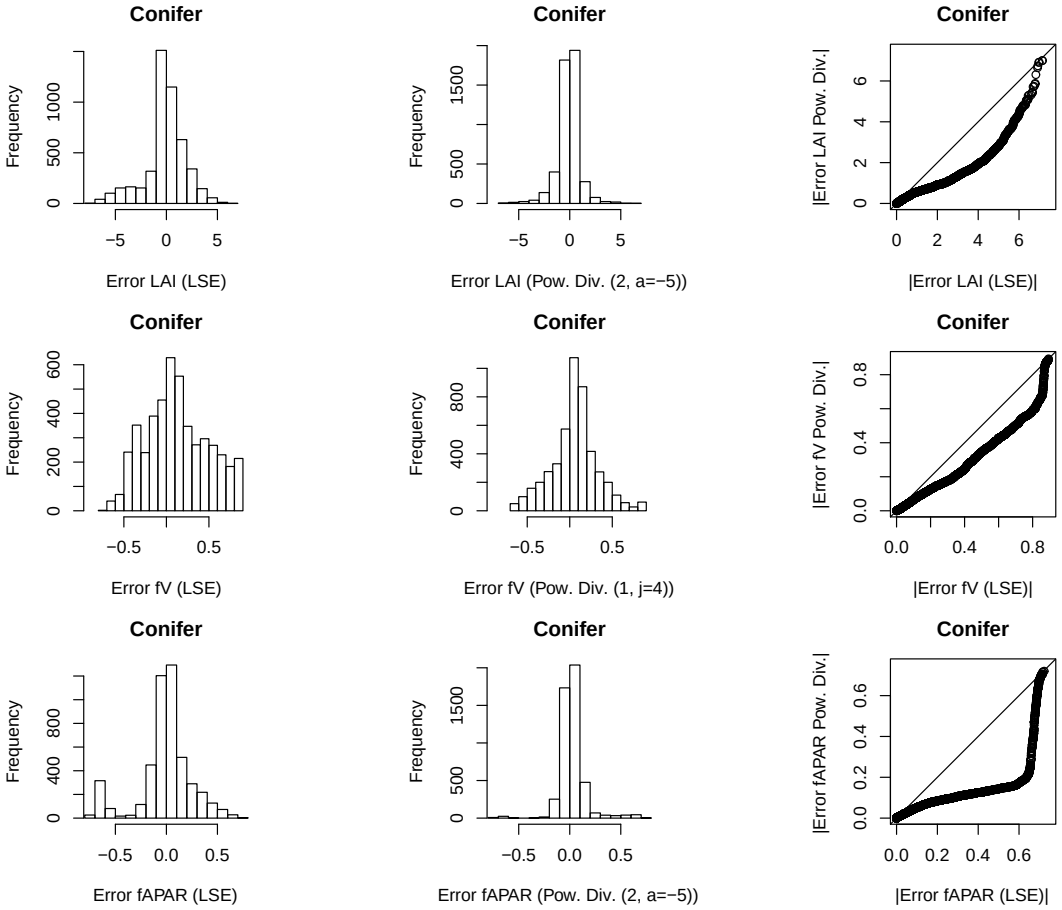


Figure 2 provides a further illustration of the reduction in the error in biophysical parameter estimation for needleleaf forests from BRF reflectance data obtained by power divergences in Equations (12) and (13) (see also Table 5).

Figure 2. Comparison of residual errors resulting from the application of two different distance measures used to estimate biophysical parameters for needleleaf canopies with a look-up table (LUT). Biophysical parameters, including leaf area index, LAI, fraction of photosynthetically active radiation absorbed by the canopy, f_{APAR}, and vegetation cover fraction, f_V, are estimated from simulated ground reflectance values. The rows from top to bottom show the comparison of residual errors in estimating LAI with the least squares estimate (LSE) and power Equation (13) with power 2 and $\alpha = -5$ (top row), of residual errors in f_V with LSE and the power divergence Equation (12) with power 1 and $j = 4$ (centre row), and of residual errors in f_{APAR} with LSE and the power Equation (13) with power 2 and $\alpha = -5$ (bottom row). The columns from left to right show frequency distributions of residual errors using LSE as a cost function (left column) and the residual errors using either a power equation for LAI and f_{APAR} or a power divergence Equation((12)) with power 1 and $j = 4$ for f_V (centre column). The right column shows the quantile–quantile plots comparing the frequency distributions of the left and centre columns expressed as absolute errors. The errors in biophysical parameters associated with the alternative cost functions are smaller and better behaved (more symmetrical and smaller bias).



The results summarized in Tables 4 and 5 and Figures 1 and 2 illustrate that the use of alternative distances measures can significantly improve parameter estimation. We found that the following distance measures perform well for the cases described above:

(1) For the broadleaf forest the best distances were:

- for the estimation of LAI: Hellinger (Equation (10)) and Arimoto (Equation (20));
- for the estimation of f_V: power divergence (Equation (12));
- for the estimation of f_{APAR}: generalized Hellinger measure (Equation (10)).

We found that compared with LSE, the Koenker–Basset distances (Equations (26) and (35)) gave better results in all cases. The improvement compared with LSE was 15% for LAI, 10% for f_V and 11% for f_{APAR}.

(2) For conifer forest the best distances were:

- for the estimation of LAI and f_{APAR}: power divergence (Equation (13)) and Pearson chi-square divergence (Equation (8));
- for the estimation of f_V: power divergence (Equation (12));

Similar to the broadleaf case, we found that the Koenker–Basset metric (Equations (26) and (35)) improved estimation in all cases. All biophysical parameters, LAI, f_V and f_{APAR}, improved by around 10% in this case.

8. Discussion

8.1. Recommendations for Distance Choice

When optimizing parameterized models for which the error distribution (shape and bias) is known, the user can choose an appropriate cost functions based on specific physical properties of the model and metrics. When we deal with non-parametric models or non-linear model with many parameters (such as the present case where we simulate BRF using PROSPECT, FLIGHT) it may be useful to check a range of available distances to get the optimal cost function. For problems similar to the present study we can provide the following guidance.

8.2. Considering the Shape of the Distribution

For non-symmetric error distributions the recommended cost function is Koenker–Basset (Equation (26)). Such non-symmetric error distribution may arise, for example, from undetected sub-pixel cloud in pre-processing. Based on the shape of this function, it is expected that for the parameter c for this function becomes close to one with increasing skewness of the error distribution. We expect skewed error distributions to be common for biophysical parameter estimation from satellite data, especially when there is a mismatch between the soil reflectance specified in the LUT and the "true" soil reflectance.

For symmetric error distributions we can recommend Hellinger (Equation (10)), Arimoto (Equation (20)) and power divergence (Equation (12)) and standard LSE (Equation (25)).

For heavy-tailed distribution (right) and semi-heavy-tailed distribution (left) (*i.e.*, the tail behave as negative exponent divided by power function) we can recommend non-symmetric power divergence (Equation (13)), Vajda (Equation (9)) and Pearson chi-square divergence (Equation (8)) cost functions. As can be seen from Figures 1 and 2, application of specific metrics to the biophysical parameters makes the errors more localized, removes heavy tails and makes the distributions more symmetric. This is due to the non-symmetric nature of the metrics themselves. To the best of our knowledge, these effects are usually not addressed in remote sensing inversion problems.

Special interest should be paid to the class of spectral metrics, since it represents informational distances in the spectral domain. Since informational transformation to the spectral domain usually makes our observation asymptotically independent, it is plausible that the spectral metric (Equation (35)) that corresponds to quasi-maximum likelihood provides good results. This is also consistent with the statistical theory that the maximum likelihood estimator is asymptotically optimal. For this reason we recommend to use this or a similar cost function.

8.3. Considering Properties of the Estimated Parameters

For all types of forest we found that for f_V the optimal metric is power divergence (Equation (12)). This symmetric cost function represents a Rényi type entropy that maximizes the entropy distribution using a power law behavior. This gives us a better understanding of the nature of errors in this case.

In the case when parameters of the model are linearly correlated, we observe consistency in the optimal cost function for these parameters (for example LAI and f_V and f_{APAR} in our case). Thus, we can recommend using the same cost function for linearly correlated parameters. However, this does not hold if the correlation has a more complicated nature.

9. Conclusions

Over 60 statistical distances from three major classes, *information measures, M-estimates and minimum contrast methods*, were obtained from the mathematical literature. A comprehensive list of these statistical distances was provided. The statistical distances were tested to see, if compared with LSE, they improved the estimation of biophysical parameters for needleleaf and broadleaf forests. We found that the commonly used LSE distance is not the optimal cost function for the cases studied and that better results can be obtained using alternative cost functions.

For the numerical experiments we use PROSPECT and FLIGHT to simulate "observed" reflectance values in 15 different spectral bands. We generate LUTs for a limited set of land-surface and atmospheric conditions. However for the observations we generate reflectance values for a wider range of conditions and thus introduce a mixture of errors caused by variations in angular geometry, LAI, f_V, soil reflectance and leaf angle distribution. For the biophysical parameter estimation we match the observed reflectances to the reflectances of the LUT with different cost functions. The largest sources of (biased) error, *i.e.*, the mismatch between observations and LUT, are potentially related to soils, since only a limited amount of variability associated with these variables was incorporated in the LUTs. We conclude that our analysis resembles a common problem for the estimation of biophysical parameters from satellite data, *i.e.*, one estimates biophysical parameters assuming a limited set of ground conditions. A cost function that is

based on an asymmetric, biased or heavy-tailed error distribution can therefore result in better estimates of biophysical parameters than LSE, which is based on a normal error distribution.

We found that the *information measures* from Section 3 provide better results when the BRF is normalized; see Equations (3) and (4). This result may not be valid for a smaller number of wavebands.

A caveat of the present study is that we analyzed only a limited subset of a wide range of possibilities, and for different applications it is likely that different cost functions may be more suitable. We are preparing a study where cost functions are used on real observations as opposed to simulated observations [25].

Alternative divergence measures and distances have been known in the statistical literature for some time and could find an application in many areas besides remote sensing, such as in biology, geography and geophysics.

The approach outlined in the present study can be extended to other applications that use LUT optimization, interpolation, linearization of parameter space, *etc*. It can be used in addition to or as an alternative to data training and machine learning schemes. We believe that the use of alternative statistical measures has great potential for remote sensing applications.

Acknowledgements

We thank Ana Prieto for advice and provision of LUT code. This work was funded by the NERC National Centre for Earth Observation (NCEO). Three anonymous reviewers are thanked for their constructive comments to improve the paper.

References

1. Sellers, P.; Los, S.O.; Tucker, C.J.; Justice, C.O.; Dazlich, D.A.; Collatz, G.J.; Randall, D.A. A revised land-surface parameterization (SiB2) for atmospheric GCMs. Part 2: The generation of global fields of terrestrial biophysical parameters from satellite data. *J. Clim.* **1996**, *9*, 706–737.
2. Jonckheere, I.; Fleck, S.; Nackaerts, K.; Muysa, B.; Coppin, P.; Weiss, M.; Baret, F. Review of methods for in situ leaf area index determination Part I. Theories, sensors and hemispherical photography. *Agric. For. Meteorol.* **2004**, *212*, 19–35.
3. Verhoef, W. Light scattering by leaf layers with application to canopy reflectance modeling. The SAIL model. *Remote Sens. Environ.* **1984**, *16*, 125–141.
4. Widlowski, J.-L.; Taberner, M.; Pinty, B.; Bruniquel-Pinel, V.; Disney, M.; Fernandes, R.; Gastellu-Etchegorry, J.-P.; Gobron, N.; Kuusk, A.; Lavergne, T.; *et al*. The third Radiation transfer Model Intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models. *J. Geophys. Res.* **2007**, doi: 10.1029/2006JD007821.
5. Iaquinta, J.; Pinty, B.; Privette, J.L. Inversion of a physically based bidirectional reflectance model of vegetation. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 687–698.
6. Gao, F.; Jin, Y.F.; Li, X.W.; Schaaf, C.B.; Strahler, A.H. Bidirectional NDVI and atmospherically resistant BRDF inversion for vegetation canopy. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1269–1278.

7. Qi, J.; Kerr, H.; Moran, M.S.; Weltz, M.; Hueye, A.R.; Dorooshian, S.; Bryant, R. Leaf area index estimates using remotely sensed data and BRDF models in a semiarid region. *Remote Sens. Environ.* **2003**, *73*, 18–30.
8. Pinty, B.; Verstraete, M.M.; Dickinson, R.E. Physical model of the bi-directional reflectance of vegetation canopies. Inversion and validation. *J. Geophys. Res.* **1990**, *95*, 11767–11775.
9. Prieto-Blanco, A.; North, P.R.J.; Barnsley, M.J.; Fox, N. Satellite-driven modelling of Net Primary Productivity (NPP): Theoretical analysis. *Remote Sens. Environ.* **2009**, *113*, 137–147.
10. Strahler, A.H. Vegetation canopy reflectance modelling. Recent developments and remote sensing perspectives. *Remote Sens. Rev*. **1997**, *15*, 179–194.
11. Qiu, J.; Gao, W.; Lesht, B.M. Inverting optical reflectance to estimate surface properties of vegetation canopies. *Int. J. Remote Sens.* **1998**, *19*, 641–656.
12. Verrelst, J.; Monoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camp-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139.
13. Richter, K.; Hank, T.B.; Vuolo, F.; Mauser, W.; D'Urso, G. Optimal exploitation of the sentinel-2 spectral capabilities for crop leaf area index mapping. *Remote Sens.* **2012**, *4*, 561–582.
14. Gascon, F.; Gastellu-Etchegorry, J.-P.; Lefevre-Fonollosa, M.-J.; Dufrene, E. Retrieval of forest biophysical variables by inverting a 3-D radiative transfer model and using high and very high resolution imagery. *Int. J. Remote Sens*. **2004**, *25*, 5601–5616.
15. North, P.R.J. Three-dimensional forest light interaction model using a Monte Carlo method. *IEEE Trans. Geosci. Remote Sens*. **1996**, *34*, 946–956.
16. Dawson, T.P.; Curran, P.J.; North, P.R.J.; Plummer, S.E. The propagation of foliar biochemical absorption features in forest canopy reflectance: A theoretical analysis. *Remote Sens. Environ.* **1999**, *67*, 147–159.
17. Privette, J.L.; Myneni, R.B.; Emery, W.J.; Pinty, B. Inversion of a soil bidirectional reflectance model for use with vegetation reflectance models. *J. Geophys. Res.* **1995**, *100*, 497–525.
18. Pardo, L. Statistical Inference Based on Divergence Measures. In *Statistics: A Series Textbooks and Monographs*; Chapman and Hall/CRC: New York, NY, USA, 2006.
19. Staudte, R.G.; Sheather, S.J. *Robust Estimation and Testing*; John Wiley & Sons, Inc.: New York, NY, USA, 1990.
20. Taniguchi, M. On estimation of parameters on Gaussian stationary processes. *J. Appl. Prob*. **1981**, *16*, 575–591.
21. Taniguchi, M. Minimum contrast estimation for spectral densities of stationary processes. *J. R. Stat. Soc.* **1987**, *9*, 315–325.
22. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat*. **1951**, *22*, 79–86.
23. Kûs, V. Blended divergences with examples. *Kybernetika* **2003**, *39*, 43–54.
24. Salicrú, M.; Menéndez, M.L.; Morales, D.; Pardo, L. Asymptotic distribution of -phi-entropies. *Commun. Statist. Theor. Method.* **1993**, *22*, 2015–2031.
25. Leonenko, G.; Los, S.O.; North, P.R.J. Retrieval of leaf area index from MODIS surface reflectance by model inversion using different minimization criteria. Unpublished work, 2013.
26. Shannon, C.E. The mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

27. Mahalanobis, P.C. On the generalised distance in statistics. *Proc. Natl. Acad. Sci. India* **1936**, *2*, 49–55.
28. Granger, C.; Lin, J.L. Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Ser. Anal.* **1994**, *15*, 371–384.
29. Rényi, A. *Probability Theory*; North-Holland: London, UK, 1970.
30. Csiszar, I. Eine informationstheoretische ungleichung and ihre anwendung auf den beweis der ergodizitat von markoffischen ketten. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **1963**, *8*, 85–108.
31. Kapur, J.N. *Maximum-Entropy Models in Science and Engineering*; John Wiley & Sons, Inc.: New York, NY, USA, 1989.
32. Lindsay, B.G. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114.
33. Menéndez, M.L.; Morales, D.; Pardo, L.; Salicrú, M. Asymptotic behavior and statistical approach of divergence measures in multinomial populations: A unified study. *Stat. Pap.* **1995**, *36*, 1–29.
34. Pardo, L.; Morales, D.; Salicrú, M.; Menéndez, M.L. Statistics in applied categorical data analysis with stratified sampling. *Utilitas Math.* **1993**, *44*, 145–164.
35. Huber, P.J. *Robust Statistics*; Wiley: New York, NY, USA, 1981.
36. Hampel, F.R.; Rousseeuw, P.J.; Ronchetti, E.M.; Stahel, W. *Robust Statistics. The Approach Based in Influence Functions*; Wiley: New York, NY, USA, 1989.
37. Edlund, O.; Ekblom, H. Computing the constrained M-estimates for regression. *Comput. Stat. Data An.* **2005**, *49*, 19–32.
38. Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50.
39. Brillinger, D.R. *The Series Data Analysis and Theory*; Hoden Day: San Francisco, CA, USA, 1981.
40. Heyde, C.C. *Quasi-Likelihood and Its Application. A General Approach to Optimal Parameter Estimation*; Springer: New York, NY, USA, 1997.
41. Dacunha-Castelle, D.; Duflo, M. *Probability and Statistics*; Springer: New York, NY, USA, 1986.
42. Liese, F.; Vajda, I. *Convex Statistical Distances*; Teubner: Leipzig, Germany, 1987.
43. Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math.* **1967**, *7*, 200–217.
44. Basu, A.; Park, C.; Lindsay, B.G.; Li, H. Some variants of minimum disparity estimation. *Comput. Stat. Data Anal.* **2004**, *45*, 741–763.
45. Read, T.R.C.; Cressie, N.A.C. *Goodness of Fit Statistics for Discrete Multivariate Data*; Springer-Verlag: New York, NY, USA, 1988.
46. Sharma, B.D.; Mittal, D.P. New non-additive measures of relative information. *J. Combinatorics Inf. Sys Sci.* **1977**, *2*, 122–133.
47. Itakura, F.; Saito, S. Analysis Synthesis Telephony Based on the Maximum Likelihood Method. In Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, 21–28 August 1968; pp. C17–C20.

48. Rényi, A. On Measures of Entropy and Information. In Proceedings of the International 4th Berkeley Symposium Mathematical Statistics Probability, Berkeley, CA, USA, 30 June–30 July 1961; Volume 1, pp. 547–561.
49. Jacquemoud, S.; Ustin, S.L.; Verdebout, J.; Schmuck, G.; Andreoli, G.; Hosgood, B. Estimating leaf biochemistry using the PROSPECT leaf optical properties model. *Remote Sens. Environ*. **1996**, *56*, 194–202.
50. Disney, M.I.; Lewis, P.; North, P.R.J. Monte Carlo ray tracing in optical canopy reflectance modelling. *Remote Sens. Rev*. **2000**, *18*, 197–226.
51. Barton, C.V.M.; North, P.R.J. Remote sensing of canopy light use efficiency using the photochemical reflectance index: Model and sensitivity analysis. *Remote Sens. Environ*. **2001**, *78*, 164–273.
52. North, P.R.J.; Rosette, J.A.B.; Suárez, J.C.; Los, S.O. A Monte Carlo radiative transfer model of satellite waveform LiDAR. *Int. J. Remote Sens*. **2010**, *31*, 1343–1358.
53. Widlowski, J.-L.; Pinty, B.; Disney, M.; Gastellu-Etchegorry, J.-P.; Lavergne, T.; Lewis, P.E.; North, P.R.J.; Pinty, B.; Thompson, R.; Verstraete, M.M. The RAMI on-line model checker (ROMC): A web-based benchmarking facility for canopy reflectance models. *Remote Sens. Environ.* **2008**, *112*, 1144–1150.
54. Thenkabail, P.S.; Enclona, E.A.; Ashton, M.S.; van Der, M. Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote Sens. Environ.* **2004**, *91*, 354–376.
55. Newcomer, J.; Landis, D.; Conrad, S.; Curd, S.; Huemmrich, K.; Knapp, D.; Morrell, A.; Nickeson, J.; Rinker, P.A.D.; Strub, R.; *et al*. *Collected Data for the Boreas Ecosystem-Atmosphere Study*; BOREAS Information System NASA/Goddard Space Flight Center: Greenbelt, MD, USA, 2000; (CD-ROM).
56. Sellers, P.; Hall, F.; Kelly, R.; Black, A.; Baldocchi, D.; Berry, J.; Ryan, M.; Ranson, J.; Crill, K.; Lettenmaier, D.; *et al*. BOREAS. Experiment overview, scientific results and future directions. *J. Geophys. Res.* **1997**, *102*, 28731–28770.
57. Gamon, J.; Huemmrich, K.; Peddle, D.; Chen, J.; Fuentes, D.; Hall, F.; Kimball, J.S.; Goetz, S.; Gu, J.; McDonald, K.C.; *et al*. Remote sensing in Boreas: Lessons learned. *Remote Sens. Environ*. **2004**, *89*, 139–162.
58. Verrelst, J.; Schaepman, M.E.; Malenovsky, Z.; Clevers, J.G.P.W. Effects of woody elements on simulated canopy reflectance: Implications for forest chlorophyll content retrieval. *Remote Sens. Environ.* **2010**, *114*, 647–656.
59. Kempeneers, P.; Zarco-Tejada, P.J.; North, P.R.J.; De Backer, S.; Delalieux, S.; Sepulcre-Cant, G.; Morales, F.; Van Aardt, J.A.N.; Sagardoy, R.; Coppin, P.; Scheunders, P. Model inversion for chlorophyll estimation in open canopies from hyperspectralimagery. *Int. J. Remote Sens*. **2008**, *29*, 5093–5111.
60. Rock, J.; Puettmann, K.J.; Gockel, H.A.; Schulte, A. Spatial aspects of the influence of silver birch (*Betula pendula* L.) on growth and quality of young oaks (*Quercus spp*.) in central Germany. *Forestry* **2004**, *77*, 235–247.

61. Grotem, R.; Reiter, I.M. Competition-dependent modelling of foliage biomass in forest stands. *Trees* **2004**, *18*, 596–607.
62. Yang, Z.; Shi, R. Calculation of Mesophyll Structure Parameter and Its Effect on Leaf Spectral Reflectance. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS '05), Seoul, Korea, 25–29 July 2005; pp. 1299–1301.
63. Baumgardner, M.F.; LeRoy, F.S.; Biehl, L.L.; Stoner, E.R. Reflectance properties of soils. *Adv. Agron.* **1985**, *38*, 1–42.
64. North, P.R.J. Estimation of f, LAI, and vegetation fractional cover from ATSR-2 imagery. *Remote Sens. Environ*. **2002**, *80*, 114–121.
65. Moses, W.J.; Bowles, J.H.; Lucke, R.L.; Corson, M.R. Impact of signal-to-noise ratio in a hyperspectral sensor on the accuracy of biophysical parameter estimation in case II waters. *Opt. Express*. **2012**, *20*, 4309–4330.

Appendix

A.1. List of f-Divergence Information Measures

Additional information about the measures below can be found in [31]. List of measures with minimum at $f(1)$.

1. Let $f(x) = x^2$ and $f(1) = 1$ then

$$D_f[P,Q] = \sum_{l=1}^{n} \frac{p_l^2}{q_l} = \sum_{l=1}^{n} \frac{(p_l - q_l)^2}{q_l} \tag{A1}$$

   This measure is called the Neyman chi-square divergence. It was introduced to test for goodness of fit in the case of multinomial probabilities. It has wide application in medical statistics.

2. Let $f(x) = (x-1)\ln(x)$ and $f(1) = 0$, then

$$D_f[P,Q] = \sum_{l=1}^{n} (p_l - q_l)\left(\ln(p_l) - \ln(q_l)\right) \tag{A2}$$

   This measure is called Jeffreys–Kullback–Leibler. It was introduced to obtain a symmetrical KL divergence.

3. K-divergence of Lin is used to analyze of contingency tables. It corresponds to the function $f(x) = x\ln(x) - x\ln\left(\frac{1+x}{2}\right)$ and $f(1) = 0$ and takes the form

$$D_f[P,Q] = \sum_{l=1}^{n} p_l \ln\left(\frac{2p_l}{p_l + q_l}\right) \tag{A3}$$

4. L-divergence of Lin is a symmetric version of K-divergence. It corresponds to the function $f(x) = x\ln(x) - (1+x)\ln\frac{1+x}{2}$ and $f(1) = 0$, thus

$$D_f[P,Q] = \sum_{l=1}^{n} p_l \ln(p_l) + q_l \ln(q_l) - (p_l + q_l)\ln\left(\frac{p_l + q_l}{2}\right) \tag{A4}$$

5. Let $f(x) = x^{\alpha}$ and $f(1) = 1$ and $x > 0$ then

$$\Gamma_f[P,Q] = \sum_{l=1}^{n} q_l \left(\frac{p_l}{q_l}\right)^{\alpha} = \sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha} \tag{A5}$$

Thus,

$$D_f[P,Q] = \sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha} - 1 \geq 0$$

is a measure of discrepancy when $\alpha > 1$. Havrada and Charvat suggested the first non-additive measure of entropy for $0 < \alpha < 1$

$$D_f[P,Q] = \frac{\sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha}}{e^{\alpha-1} - 1} \tag{A6}$$

For $\alpha \to 1$ it turns to be Kullback–Leibler's distance. When $0 < \alpha < 1$, $\sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha}$ is positive concave function and so its logarithm is also a concave function.

6. We can use a higher order cross-entropy, or the so-called cross-entropy of order α

$$D_f[P,Q] = \frac{1}{\alpha - 1} \ln \left(\sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha} \right), \ \alpha > 0, \ \alpha \neq 1 \tag{A7}$$

This measure was suggested by Rényi [29] and it plays an important role in ecology and statistics as indices of diversity. It is related to the Shannon entropy of integer order.

7. Liese and Vajda [42] proposed bounded asymptotic Rényi measure, which has application in signal detection problem. For all $\alpha \neq 0, 1$ we define

$$D_f[P,Q] = \frac{1}{\alpha(\alpha - 1)} \ln \left(\sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha} \right) \tag{A8}$$

8. The harmonic Toussaint measure corresponds to the function $f(x) = x\frac{x-1}{x+1}$ and $f(1) = 0$. It has application in measuring the dissimilarity between musical rhythms, music information retrieval and copyright infringement resolution to computational music theory and evolutionary studies of music. The measure has the following form

$$D_f[P,Q] = \sum_{l=1}^{n} \left(p_l - \frac{2p_l q_l}{p_l + q_l} \right) \tag{A9}$$

Distances with the minimum at zero:

1. The negative exponential disparity measure is used as an estimator that is asymptotically fully efficient and is robust against outliers and inliers, see [32]:

$$D_f[P,Q] = \sum_{l=1}^{n} q_l (exp(-\frac{p_l - q_l}{q_l}) - 1) \tag{A10}$$

2. The Bregman divergences [43] are not full distance measures, because they does not satisfy the triangle inequality and they are not symmetric. Bregman divergences are important for two

reasons: Firstly, they generalize squared Euclidean distances to a class of distances that all share similar properties. Secondly, they bear a strong connection to exponential families of distributions.

$$D_f[P,Q] = \sum_{l=1}^{n}(q_l^a + \frac{1}{a-1}p_l^a - \frac{1}{a-1}p_l q_l^{a-1}),\ a \neq 1 \tag{A11}$$

The symmetrized version of the Bregman divergences can be presented in a form

$$D_f[P,Q] = \sum_{l=1}^{n}(p_l - q_l)[p_l^{a-1} - q_l^{a-1}]$$

3. The powered Pearson divergence is reasonably efficient and robust [44], and has wide applicability in genetics.

$$D_f[P,Q] = \frac{1}{2\alpha^2}\sum_{l=1}^{n} q_l \left(\frac{\{p_l{}^{\alpha} - q_l^{\alpha}\}}{q_l^{\alpha}}\right)^2,\ \alpha \in (0,1] \tag{A12}$$

It includes the Pearson's chi-squared measure $(\alpha = 1)$ and the Hellinger's measure $(\alpha = 1/2)$.

4. The Cressie and Read power divergence [45] results in stable disparity measures and solutions when outliers are added to the data. Under certain general conditions this estimator has asymptotic breakdown points of 50%.

$$D_f[P,Q] = \frac{1}{\alpha(\alpha+1)}\left[\sum_{l=1}^{n} p_l[p_l/q_l]^{\alpha} - 1\right],\ -\infty < \alpha < \infty \tag{A13}$$

5. The Sharma and Mittal divergences [46] are two generalizations of the Kullback–Leibler measure. One is called α-order and β-degree divergence measure and the other is called 1-order and β-degree divergence measure

$$D_f[P,Q] = \frac{1}{(\beta-1)}\left[\left(\sum_{l=1}^{n} p_l^{\alpha} q_l^{1-\alpha}\right)^{\frac{\beta-1}{\alpha-1}} - 1\right],\ \alpha,\beta \neq 1 \tag{A14}$$

$$D_f[P,Q] = \frac{1}{(\beta-1)}\left[exp\left((\beta-1)\sum_{l=1}^{n} p_l log\left(\frac{p_l}{q_l}\right)\right) - 1\right],\ \beta \neq 1 \tag{A15}$$

6. Finally, the β-divergence [47], is given by

$$D_f[P,Q] = \begin{cases} \sum_{l=1}^{n}\left(\frac{1}{\beta(\beta-1)}\right)\left(p_l^{\beta} + (\beta-1)q_l^{\beta} - \beta p_l q_l^{\beta-1}\right) p_l/q_l - log[p_l/q_l] - 1, & \beta \in \mathbb{R}/(0,1) \\ \sum_{l=1}^{n} p_l(log[p_l] - log[q_l]) + (q_l - p_l), & \beta = 1 \\ \sum_{l=1}^{n} p_l/q_l - log[p_l/q_l] - 1, & \beta = 0 \end{cases} \tag{A16}$$

This divergence was introduced by Itakura–Saito for the estimation of short-time speech spectra using an autoregressive model. It became popular in speech and acoustics research and it was applied to denoising and up-mix (mono to stereo conversion) of music.

A.2. List of the Measures between Two Probability Distributions

List of (h, f)-measures [18]:

1. Sharma–Mittal divergence [46] with

$$h(x) = \frac{1}{(\beta-1)}\left((1+\alpha(\alpha-1)x)^{\frac{\beta-1}{\alpha-1}}-1\right);\ f(x) = \frac{x^{\alpha}-\alpha(x-1)-1}{\alpha(\alpha-1)};\ \alpha \neq 0,1; \quad \text{(A17)}$$

$$D_f^h[P,Q] = \frac{1}{(\beta-1)}\left(\left[1+\sum_{l=1}^{n} q_l(p_l/q_l)^{\alpha}-\alpha(p_l-q_l)-q_l\right]^{\frac{\beta-1}{\alpha-1}}-1\right)$$

2. Bhattacharyya divergence has interesting application in signal selection and it has the following form

$$h(x) = -log(-x+1);\ f(x) = -x^{1/2}+(1/2)(x+1) \quad \text{(A18)}$$

$$D_f^h[P,Q] = -log\left(1+\sum_{l=1}^{n}\sqrt{p_l q_l}-\frac{1}{2}(p_l+q_l)\right)$$

List of the entropy measures:

1. Shannon (1948) [26]

$$f(x) = -xlogx,\ h(x) = x$$

$$D_f^h(P,Q) = -\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)log\left(\frac{p_l+q_l}{2}\right)+\frac{1}{2}\left(\sum_{l=1}^{n}p_l log(p_l)+\sum_{l=1}^{n}q_l log(q_l)\right) \quad \text{(A19)}$$

2. Rényi (1961) [48]

$$f(x) = x^{\alpha},\ h(x) = \left[\frac{1}{\alpha(1-\alpha)}logx\right],\ \alpha \neq 0,1$$

$$D_f^h(P,Q) = \left(\frac{1}{\alpha(1-\alpha)}\right)\left[log\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha}\right)-\frac{1}{2}\left(log\sum(p_l)^{\alpha}+log\sum_{l=1}^{n}(q_l)^{\alpha}\right)\right] \quad \text{(A20)}$$

3. Varma (1966) [18]

$$f(x) = x^{\alpha-\beta+1},\ h(x) = \left[\frac{1}{\beta-\alpha}logx\right],\ \beta-1<\alpha<\beta,\ \beta \geqslant 1$$

$$D_f^h(P,Q) = \left(\frac{1}{\beta-\alpha}\right)\left[log\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha-\beta+1}\right)-\frac{1}{2}\left(log\sum(p_l)^{\alpha-\beta+1}+log\sum_{l=1}^{n}(q_l)^{\alpha-\beta+1}\right)\right] \quad \text{(A21)}$$

4. Varma (1966) [18]

$$f(x) = x^{\alpha/\beta},\ h(x) = \left[\frac{1}{\beta(\beta-\alpha)} logx\right],\ 0 < \alpha < \beta,\ \beta \geqslant 1$$

$$D_f^h(P,Q) = \left(\frac{1}{\beta(\beta-\alpha)}\right)\left[log\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha/\beta}\right) - \frac{1}{2}\left(log\sum(p_l)^{\alpha/\beta} + log\sum_{l=1}^{n}(q_l)^{\alpha/\beta}\right)\right] \tag{A22}$$

5. Havrda and Charvat (1967) [18]

$$f(x) = \frac{1}{1-\alpha}(x^\alpha - x),\ h(x) = x,\ \alpha > 0,\ \alpha \neq 1$$

$$D_f^h(P,Q) = \left(\frac{1}{1-\alpha}\right)\left[\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha} - \left(\frac{p_l+q_l}{2}\right)\right) - \frac{1}{2}\left(\sum({p_l}^\alpha - p_l) + \sum({q_l}^\alpha - q_l)\right)\right] \tag{A23}$$

6. Sharma and Mittal [46]

$$f(x) = xlogx,\ h(x) = \frac{exp((\alpha-1)x)-1}{(1-\alpha)},\ \alpha > 0,\ \alpha \neq 1$$

$$Df(P,Q) = \left(\frac{1}{1-\alpha}\right) exp\left((\alpha-1)\sum_{l=1}^{n}\left(\left(\frac{p_l+q_l}{2}\right) log\left(\frac{p_l+q_l}{2}\right)\right)\right) \tag{A24}$$

$$-\left(\frac{1}{1-\alpha}\right)\left[\frac{1}{2}\left(exp\left((\alpha-1)\sum_{l=1}^{n}(p_l logp_l)\right) + exp\left((\alpha-1)\sum_{l=1}^{n}(q_l logq_l)\right)\right)\right]$$

7. Sharma and Mittal [46]

$$f(x) = x^\beta,\ h(x) = \frac{1}{(1-\alpha)}\left(x^{\frac{\alpha-1}{\beta-1}} - 1\right)\ \alpha > 0,\ \beta > 0\ \alpha, \beta \neq 1$$

$$D_f^h(P,Q) = \left(\frac{1}{1-\alpha}\right)\left[\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\beta}\right)^{\frac{\alpha-1}{\beta-1}} - \frac{1}{2}\left(\left[\sum_{l=1}^{n}({p_l}^\beta)\right]^{\frac{\alpha-1}{\beta-1}} + \left[\sum_{l=1}^{n}({q_l}^\beta)\right]^{\frac{\alpha-1}{\beta-1}}\right)\right] \tag{A25}$$

8. Ferreri (1980) [18]

$$f(x) = (1+\alpha x)log(1+\alpha x),\ h(x) = \left(1+\frac{1}{\alpha}\right) log(1+\alpha) - \frac{x}{\alpha},\ \alpha > 0$$

$$D_f^h(P,Q) = -\frac{1}{\alpha}\left(\sum_{l=1}^{n}\left(1+\alpha\left[\frac{p_l+q_l}{2}\right]\right) log\left(1+\alpha\left[\frac{p_l+q_l}{2}\right]\right)\right) \tag{A26}$$

$$+\frac{1}{\alpha}\left[\frac{1}{2}\left(\sum_{l=1}^{n}(1+\alpha p_l)\,log\,(1+\alpha p_l) + \sum_{l=1}^{n}(1+\alpha q_l)\,log\,(1+\alpha q_l)\right)\right]$$

9. Kapur (1972) [18]

$$f(x) = \frac{x^{\alpha} + (1-x)^{\alpha} - 1}{(1-\alpha)},\ h(x) = x,\ \alpha \neq 1$$

$$D_f^h(P,Q) = \left(\frac{1}{1-\alpha}\right)\left(\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha} + \left(1-\left(\frac{p_l+q_l}{2}\right)\right)^{\alpha} - 1\right) \quad \text{(A27)}$$

$$-\left(\frac{1}{1-\alpha}\right)\left[\frac{1}{2}\left(\left[\sum_{l=1}^{n} p^{\alpha} + (1-p_l)^{\alpha} - 1\right] + \left[\sum_{l=1}^{n} q^{\alpha} + (1-q_l)^{\alpha} - 1\right]\right)\right]$$

10. Burbea (1984) [18]

$$f(x) = \frac{x^{\alpha} - (1-x)^{\alpha} + 1 + (\alpha-1)^{-1}(2^{\alpha}-2)x}{(\alpha-2)},\ h(x) = x,\ \alpha > 0,\ \alpha \neq 1$$

$$D_f^h(P,Q) = \left(\frac{1}{\alpha-2}\right)\left[\sum_{l=1}^{n}\left(\frac{p_l+q_l}{2}\right)^{\alpha} - \left(1-\left(\frac{p_l+q_l}{2}\right)\right)^{\alpha} + 1 + (\alpha-1)^{-1}(2^{\alpha}-2)\left(\frac{p_l+q_l}{2}\right)\right] \quad \text{(A28)}$$

$$-\frac{1}{2}\left(\frac{1}{\alpha-2}\right)\left[\sum_{l=1}^{n} p^{\alpha} - (1-p_l)^{\alpha} + 1 + (\alpha-1)^{-1}(2^{\alpha}-2)p_l\right]$$

$$-\frac{1}{2}\left(\frac{1}{\alpha-2}\right)\left[\sum_{l=1}^{n} q^{\alpha} - (1-q_l)^{\alpha} + 1 + (\alpha-1)^{-1}(2^{\alpha}-2)q_l\right]$$

A.3. List of the Blended f-Disparities

More information of the measures below could be found in [23].

1. Pearson–Neyman blend with corresponding blended divergence

$$D_{\beta}(P,Q) := \frac{1}{2}\sum_{l=1}^{n}\frac{(p_l-q_l)^2}{\beta p_l + (1-\beta)q_l} \quad \text{(A29)}$$

This blend coincides with the generalized LeCam divergence. Also it is bounded and

$$0 \leq D_f(P,Q) \leq \frac{1}{2}\frac{1}{1-\beta} + \frac{1}{2\beta}$$

2. Blended power divergence-variant A. For $a \in R - \{0,1\}$ we have

$$D_{a,\beta}(P,Q) = \frac{1}{a(a-1)}\sum_{l=1}^{n}\frac{p_l^a + q_l^a}{(\beta p_l + (1-\beta)q_l)^{a-1}} - 2,\ a \neq 0,1 \quad \text{(A30)}$$

Note that $D_{0,\beta}(P,Q)$ are unbounded for all $\beta \in [0,1)$, which corresponds to the reversed Kullback blend. And $D_{1,\beta}(P,Q)$ for all $\beta \in (0,1)$- Kullback-reversed blend is bounded and

$$0 \leq D_f(P,Q) \leq -\ln(1-\beta) - \ln\beta$$

$$D_{1,\beta}(P,Q) = \sum_{l=1}^{n} p_l \ln\left(\frac{p_l}{\beta p_l + (1-\beta)q_l}\right) + q_l \ln\left(\frac{q_l}{\beta p_l + (1-\beta)q_l}\right) \quad \text{(A31)}$$

3. Blended power divergence-variant B. For $0 < |a| < 1$ and $\beta \in (0, 1)$

$$Da, \beta(P, Q) = -sign(a) \sum_{l=1}^{n} \frac{(\beta p_l + (1-\beta) q_l) p_l^a + q_l^{a+1}}{(\beta p_l + (1-\beta) q_l)^a}, \ 0 < |a| < 1 \tag{A32}$$

is bounded, but not symmetric.

A.4. List of M-Estimates

Information of the measures below is based on [35–37].

1. More general estimates with Laplace distribution

$$\rho(x) = |x|^p, \psi(x) = psign(x) |x|^{p-1}, 1 \leq p \leq 2 \tag{A33}$$

are known as L_p-estimates. It has been used in statistics of speech and image data processing, especially when observed in a transform domain like the wavelet or discrete Fourier transform domains. For example, the over complete wavelet transform coefficients of images are found to have sparse distributions, a property that has been extensively exploited in coding and denoising. It appears that p must be fairly moderate to provide a relatively robust estimator or, in other words, to provide an estimator scarcely perturbed by outlying data.
2. For positive α, the function

$$\rho(x) = \frac{1}{2\alpha} - \frac{exp(-\alpha x^2)}{2\alpha}, \ \psi(x) = x \exp\{-\alpha x^2\} \tag{A34}$$

leads to the so-called alpha estimator. It has been applied in a special class of change-point models, where the change is defined as a shift of observations means.
3. For $\nu, s > 0$, we can determine the trigonometric and the hyperbolic estimators

$$\rho(x) = \nu \left(xarctan(sx) - \frac{log(s^2 x^2 + 1)}{2s} \right), \psi(x) = \nu \arctan(sx) \tag{A35}$$

$$\rho(x) = \nu \frac{log(cosh(sx))}{s}, \ \psi(x) = \nu \tanh(sx) \tag{A36}$$

Errors in this case are distributed by a logistic distribution.
4. For the Cauchy distribution $f(x) = 1/(\pi(1 + x^2))$ we have

$$\rho(x) = \frac{c2}{2} log((x^2/c^2) + 1), \ \psi(x) = \frac{2x}{1 + x^2/c^2} \tag{A37}$$

Cauchy's function, also known as the Lorentzian function, does not guarantee a unique solution (unicity). With a descending first derivative, such a function has a tendency to yield erroneous solutions. It has been applied in image analysis and in particular to the problem of parametric image registration.
5. Latter influence functions trimmed at $0 < c < \infty$ are presented in the form

$$\psi(x) = x 1_{(-c,c)}(x), \qquad \psi(x) = 1_{(-c,c)}(x) sign(x) \tag{A38}$$

and they specify the trimmed least squares and the trimmed absolute error estimators.

6. The Welsh distance has the form

$$\rho(x) = \frac{c^2}{2}(1 - exp(-x^2/c^2)) \tag{A39}$$

As can be seen from the influence function, the influence of large errors only decreases linearly with their size. However it has the same problem as the Cauchy distance.

7. The Geman and McClure function tries to reduce the effect of large errors further, but it also cannot guarantee unicity.

$$\rho(x) = x^2(1 + x^2)^{-1} \tag{A40}$$

It has been applied successfully in optical flow estimation, image restoration and vision-based recognition in continuous dynamic hand gestures.

8. The Tukey function also encounters the problem of unicity; it can be written as

$$\rho(x) = \begin{cases} \frac{c^2}{6}(3x^2 - 3x^4 + x^6), & |x| \leq c \\ c^2/6 & |x| > c \end{cases} \tag{A41}$$

It has applications in economics, computer vision and satellite retrievals.

9. The Huber function has the following form

$$\rho_c(x) = \begin{cases} \frac{1}{2}x^2, & |x| < c \\ c\,|x| - \frac{1}{2}c^2, & |x| \geq c \end{cases} \tag{A42}$$

and the score function is represented as

$$\psi_c(x) = \max\{\min(x, c), -c\}, c > 0$$

It has some optimal properties, see [35], and the function $\psi_c(x)$ can be approximated by a twice differentiable score function. Huber's M-estimation has been applied in GPS positioning and modeling of complex technical experiments where it reduces the effect of outliers. This estimator is so satisfactory that it has been recommended for almost all situations, however, from time to time, difficulties are encountered, which may be related to a lack of stability.

A.5. Minimum Contrast Estimation List

Detailed information of the following measures can be found in [20,21].

1. let $K(x) = -logx + x$, then

$$L(f_\theta, g) = \sum_{\lambda_j \in \Lambda} \{-log(f_\theta(\lambda_j)/g(\lambda_j)) + f_\theta(\lambda_j)/g(\lambda_j)\} \tag{A43}$$

2. let $K(x) = (logx)^2$, then

$$L(f_\theta, g) = \sum_{\lambda_j \in \Lambda} \{logf_\theta(\lambda_j) - log(g(\lambda_j))\}^2 \tag{A44}$$

3. let $K(x) = xlogx - x$, then

$$L(f_\theta, g) = \sum_{\lambda_j \in \Lambda} f_\theta(\lambda_j)g(\lambda_j)^{-1}\{log(f_\theta(\lambda_j)g(\lambda_j)^{-1}) - 1\} \tag{A45}$$

4. let $K(x) = (x^{\alpha} - 1)^2$, where $0 < \alpha < \infty$ then

$$L(f_{\theta}, g) = \sum_{\lambda_j \in \Lambda} \{(f_{\theta}(\lambda_j)/g(\lambda_j))^{\alpha} - 1\}^2 \tag{A46}$$

This is an α-entropy criterion for a Gaussian process.