mathematics

MDPI

Article

Qualitative Properties of Randomized Maximum Entropy Estimates of Probability Density Functions

Yuri S. Popkov [1,2,3]

[1] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 119333 Moscow, Russia; popkov@isa.ru
[2] Institute of Control Sciences of Russian Academy of Sciences, 117997 Moscow, Russia
[3] Department of Software Engineering, ORT Braude College, 2161002 Karmiel, Israel

Abstract: The problem of randomized maximum entropy estimation for the probability density function of random model parameters with real data and measurement noises was formulated. This estimation procedure maximizes an information entropy functional on a set of integral equalities depending on the real data set. The technique of the Gâteaux derivatives is developed to solve this problem in analytical form. The probability density function estimates depend on Lagrange multipliers, which are obtained by balancing the model's output with real data. A global theorem for the implicit dependence of these Lagrange multipliers on the data sample's length is established using the rotation of homotopic vector fields. A theorem for the asymptotic efficiency of randomized maximum entropy estimate in terms of stationary Lagrange multipliers is formulated and proved. The proposed method is illustrated on the problem of forecasting of the evolution of the thermokarst lake area in Western Siberia.

Keywords: randomized maximum entropy estimation; probability density functions; Lagrange multipliers; Lyapunov-type problems; implicit function; rotation of vector field; asymptotic efficiency; thermokarst lakes; forecasting

check for updates

Citation: Popkov, Y.S. Qualitative Properties of Randomized Maximum Entropy Estimates of Probability Density Functions. *Mathematics* **2021**, *9*, 548. https://doi.org/10.3390/math9050548

Academic Editors: Mikhail Posypkin, Andrey Gorshenin and Vladimir Titarev

Received: 28 January 2021
Accepted: 2 March 2021
Published: 5 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Estimating the characteristics of models is a very popular and, at the same time, important problem of science. This problem arises in applications with unknown parameters, which have to be estimated somehow using real data sets. In particular, such problems have turned out to be fundamental in machine learning procedures [1–5]. The core of these procedures is a parametrized model trained by statistically estimating the unknown parameters based on real data. Most of the econometric problems associated with reconstructing functional relations and forecasting also reduce to estimating the model parameters; for example, see [6,7].

The problems described above are solved using traditional mathematical statistics methods, such as the maximum likelihood method and its derivatives, the method of moments, Bayesian methods, and their numerous modifications [8,9].

Among the mathematical tools for parametric estimation mentioned, a special place is occupied by entropy maximization methods for finite-dimensional probability distributions [10,11].

Consider a random variable x taking discrete values $x_1, \dots, x_n$ with probabilities $p_1, \dots, p_n$, respectively, and r functions $f_1(x), \dots, f_r(x)$ of this variable with discrete values. The discrete probability distribution function $\mathbf{p}(x) = \{p_1(x_1), \dots, p_n(x_n)\}$ is defined as the solution of the problem

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i, \quad \sum_{i=1}^{n} p_i f_k(x_i) \le q_k, \quad k = 1, \dots, r,$$

where $q_1, \ldots, q_r$ are given constants.

If $f_k(x_i) \equiv x_i^k$, then the system of equalities specifies constraints on the kth moments of the discrete random variable x. In the case of equality constraints, some modifications of this problem adapted to different applications were studied in [10–13]. Since this problem is conditionally extremal, it can be solved using the Lagrange method, which leads to a system of equations for Lagrange multipliers. The latter often turn out to be substantially nonlinear functions, and hence, rather sophisticated techniques are needed for their numerical calculation [14,15].

In the case of inequality constraints, this problem belongs to the class of mathematical programming problems [16].

The entropy maximization principle is adopted to estimate the parameters of a priori distributions when constructing Bayesian estimates [17,18] or maximum likelihood estimates.

The parameters of probability distributions (continuous or discrete) can be estimated using various mathematical statistics methods, including the method of entropy maximization. Their efficiency in hydrological problems was compared in [19]. Apparently, the method of entropy maximization yields the best results in such problems due to the structure of hydrological data.

The problem of estimating some model characteristics on real data was further developed in connection with the appearance of new machine learning methods, called randomized machine learning (RML) [20]. They are based on models with random parameters, and it is necessary to estimate the probability density functions of these parameters. The estimation algorithm (RML algorithm) is formulated in terms of functional entropy-linear programming [21].

The original statement of this problem was to estimate probability density functions (PDFs) in RML procedures. However, in recent times, a more general context has been assumed—the method of maximizing entropy functionals for constructing estimates of continuous probability density functions using real data (randomized maximum entropy (RME) estimation).

In this paper, the general RME estimation problem is formulated; its solutions, numerical algorithms, and the asymptotic properties of the solutions are studied. The theoretical results are illustrated by an important application—estimating the evolution of the thermokarst lake area in Western Siberia.

2. Statement of the RME Estimation Problem

Consider a scalar continuous function $\varphi(x, \theta)$ with parameters $\theta = \{\theta_1, \ldots, \theta_n\}$. Assume that this function is a characteristic of an object's model with an input x and an output $\hat{y}$. Let $\mathbf{x}^{(r)} = \{x[1], \ldots, x[r]\}$ and $\mathbf{y}^{(r)} = \{y[1], \ldots, y[r]\}$ be given measurements at time $t = 1, \ldots, r$. Note that the latter measurements are obtained with random vector errors $\xi = \{\xi[1], \ldots, \xi[r]\}$, which are generally different for different time points.

Thus, after r measurements, the model and observations are described by the equations

$$\begin{aligned} \hat{\mathbf{y}} &= \Gamma(\mathbf{x}^{(r)}, \theta), \\ \hat{\mathbf{v}} &= \hat{\mathbf{y}} + \xi, \end{aligned} \tag{1}$$

where the vector function $\Gamma(\mathbf{x}^{(r)}, \theta)$ has the components $\varphi(x[t], \theta)$, where $t = 1, \ldots, r$ are the time points; $\hat{\mathbf{v}}$ denotes the observed output of the model containing measurement noises of the object's output.

Let us introduce a series of assumptions necessary for further considerations.

- The random parameters are $\theta \in \Theta \subset R^n$, $\Theta = [\theta^-, \theta^+]$, where $[\bullet, \bullet]$ is a vectorial segment in the space R^n [22].
- The PDF $P(\theta)$ of the parameters is continuously differentiable on its support Θ.

- The random noise is $\xi \in \Xi \subset R^r$, where

$$\Xi = \bigotimes_{t=1}^{r} \Xi_t, \qquad \Xi_t = [\xi_t^-, \xi_t^+]. \tag{2}$$

- The PDF $Q(\xi)$ of the measurement noises is continuously differentiable on the support Ξ and also has the multiplicative structure

$$Q(\xi) = \prod_{t=1}^{r} Q_t(\xi[t]). \tag{3}$$

The estimation problem is stated as follows: Find the estimates of the PDFs $P^*(\theta)$ and $Q^*(\xi)$ that maximize the generalized information entropy functional

$$\mathcal{H}[P(\theta), Q(\xi)] = -\int_Q P(\theta) \ln P(\theta) d\theta - \sum_{t=1}^{r} \int_{\Xi_t} Q_t(\xi[t]) \ln Q_t(\xi[t]) \Rightarrow \max \tag{4}$$

subject to

—the normalization conditions of the PDFs given by

$$\int_\Theta P(\theta) d\theta = 1; \quad \int_{\Xi_t} Q_t(\xi[t]) d\xi[t] = 1, \quad t = 1, \ldots, r; \tag{5}$$

and

—the empirical balance conditions

$$\begin{aligned} \Phi[P(\theta), Q(\xi)] &= \mathbf{y}^{(r)}, \\ \Phi[P(\theta), Q(\xi)] &= \{\Phi_1[P(\theta), Q(\xi)], \ldots, \Phi_r[P(\theta), Q(\xi)]\} \\ \Phi_t[P(\theta), Q(\xi)] &= \int_\Theta \varphi(x[t], \theta) P(\theta) d\theta + \int_{\Xi_t} Q_t(\xi[t]) \xi[t] d\xi[t], \quad t = 1, \ldots, r, \end{aligned} \tag{6}$$

where $\mathbf{y}^{(r)} = \{y[1], \ldots, y[r]\}$ are the measured data on the object's output. We will denote the problems (4)–(6) as the RME estimate.

Problems (4)–(6) are of the Lyapunov type [23,24], as they have an integral functional and also integral constraints.

3. Optimality Conditions

The optimality conditions in optimization problems of the Lyapunov type are formulated in terms of Lagrange multipliers. In addition, the Gâteaux derivatives of the problem's functionals are used [25].

The Lagrange functional is defined by

$$\begin{aligned} \mathcal{L}[P(\theta), Q(\xi), \mu, \eta, \lambda] &= \mathcal{H}[P(\theta), Q(\xi)] + \mu \left(1 - \int_\Theta P(\theta) d\theta\right) + \\ &+ \sum_{t=1}^{r} \eta_t \left(1 - \int_{\Xi_t} Q_t(\xi[t]) d\xi[t]\right) + \\ &+ \sum_{t=1}^{r} \lambda_t \left(y[t] - \int_\Theta P(\theta) \varphi(x[t], \theta) d\theta - \int_{\Xi_t} Q_t(\xi[t]) \xi[t] d\xi[t]\right). \end{aligned} \tag{7}$$

Let us recall the technique for obtaining optimality conditions in terms of the Gâteaux derivatives [26].

The PDFs $P(\theta)$ and $Q_t(\xi[t])$, $(t = 1, \ldots, r)$, are continuously differentiable, i.e., belong to the class C^1. Choosing arbitrary functions $h(\theta)$ and $w_t(\xi[t])$, $(t = 1, \ldots, r)$, from this class, we represent the PDFs as

$$P(\theta) = P^*(\theta) + \alpha h(\theta); \quad Q_t(\xi[t]) = Q_t^*(\xi[t]) + \beta_t w_i(\xi[t]), \quad t = 1, \ldots, r,$$

where the PDFs $P^*(\theta)$ and $Q_t^*(\xi[t])$ are the solutions of problems (4)–(6), and α and $\beta_1, \ldots, \beta_r$ are parameters.

Next, we substitute the above representations of the PDFs into (7). If all functions from C^1 are assumed to be fixed, the Lagrange functional depends on the parameters α and $\beta_1, \ldots, \beta_r$. Then, the first-order optimality conditions for the functional (7) in terms of the Gâteaux derivative take the form

$$\left.\frac{d\mathcal{L}}{d\alpha}\right|_{(\alpha,\beta)=0} = 0, \quad \left.\frac{\partial\mathcal{L}}{\partial\beta_t}\right|_{(\alpha,\beta)=0} = 0, \quad t = 1, \ldots, r.$$

These conditions lead to the following system of integral equations:

$$\int_\Theta h(\theta)\Omega(\theta)d\theta = 0, \quad \int_{\Xi_t} w_t(\xi[t])\mathrm{Y}_t(\xi[t])d\xi[t] = 0, \quad t = 1, \ldots, r,$$

which are satisfied for any functions $h(\theta)$ and $w_1(\xi[1]), \ldots, w_r(\xi[r])$ from C^1 if and only if

$$\Omega(\theta) = 0, \quad \mathrm{Y}_t(\xi[t]) = 0, \quad t = 1, \ldots, r.$$

The optimality conditions for problems (4)–(6) are given by

$$\Omega(\theta) = \ln P^*(\theta) + 1 - \mu - \sum_{t=1}^{s} \lambda_t \varphi(x[t], \theta) = 0, \tag{8}$$

$$\mathrm{Y}_t(\xi[t]) = \ln Q_t^*(\xi[t]) + 1 - \eta_t - \lambda_t \xi[t] = 0, \quad t = 1, \ldots, r. \tag{9}$$

Hence, the entropy-optimal PDFs of the model parameters and measurement noises have the form

$$\begin{aligned} P^*(\theta \,|\, \mathbf{y}^{(r)}, \mathbf{x}^{(r)}) &= \frac{\exp\Big(-\sum_{j=1}^{r} \lambda_j(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\varphi(x[j], \theta)\Big)}{\mathcal{P}(\lambda(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})}, \\ Q_t^*(\xi[t] \,|\, \mathbf{y}^{(r)}, \mathbf{x}^{(r)}) &= \frac{\exp\Big(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\xi[t]\Big)}{\mathcal{Q}_t(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})}, \quad t = 1, \ldots, r, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \mathcal{P}(\lambda(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}) &= \int_\Theta \exp\left(-\sum_{j=1}^{r} \lambda_j(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\varphi(x[j], \theta)\right) d\theta, \\ \mathcal{Q}_t(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}) &= \int_{\Xi_t} \exp\Big(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\xi[t]\Big) d\xi[t], \quad t = 1, \ldots, r. \end{aligned} \tag{11}$$

Due to equalities (10) and (11), the entropy-optimal PDFs are parametrized by the Lagrange multipliers $\lambda_1, \ldots, \lambda_r$, which represent the solutions of the empirical balance equations

$$\frac{\mathcal{G}(\lambda(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}))}{\mathcal{P}(\lambda(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}))} + \frac{\mathcal{E}_t(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}))}{\mathcal{Q}_t(\lambda_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}))} = y[t], \quad t = 1, \ldots, r, \tag{12}$$

where

$$\mathcal{G}(\lambda(\mathbf{y}^{(r)},\mathbf{x}^{(r)})) = \int_{\Theta} \varphi(x[t],\theta) \exp\left(-\sum_{j=1}^{r} \lambda_j(\mathbf{y}^{(r)},\mathbf{x}^{(r)})\varphi(x[j],\theta)\right) d\theta,$$

$$\mathcal{E}_t(\lambda_t(\mathbf{y}^{(r)},\mathbf{x}^{(r)})) = \int_{\Xi_t} \xi[t] \exp\Big(-\lambda_t(\mathbf{y}^{(r)},\mathbf{x}^{(r)})\xi[t]\Big) d\xi[t], \quad t = 1,\dots,r. \tag{13}$$

The solution $\lambda^*(\mathbf{y}^{(r)},\mathbf{x}^{(r)})$ of these equations depends on the sample $(\mathbf{y}^{(r)},\mathbf{x}^{(r)})$ used for constructing the RME estimates of the PDFs.

4. Existence of an Implicit Function

The second term in the balance Equations (12) and (13) is the mean value of the noise in each measurement t. The noises and their characteristics are often assumed to be equal over the measurements:

$$\xi^- \le \xi[t] \le \xi^+, \quad t = 1,\dots,r. \tag{14}$$

Therefore, the mean value of the noise is given by

$$\bar{\xi} = \frac{\mathcal{E}_t(\lambda_t(\mathbf{y}^{(r)},\mathbf{x}^{(r)}))}{\mathcal{Q}_t\big(\lambda_t(\mathbf{y}^{(r)},\mathbf{x}^{(r)})\big)}, \quad \xi^- \le \bar{\xi} \le \xi^+. \tag{15}$$

The balance equations can be written as

$$\begin{aligned} W_t(\lambda \mid \tilde{y}[t],\mathbf{x}^{(r)}) &= \int_{\Theta} (\varphi(x[t],\theta) - \tilde{y}[t]) \exp\left(-\sum_{j=1}^{r} \lambda_j(\tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)})\varphi(x[j],\theta)\right) d\theta = 0, \\ t &= 1,\dots,r, \end{aligned} \tag{16}$$

where

$$\tilde{y}[t] = y[t] - \bar{\xi}, \qquad \tilde{\mathbf{y}}^{(r)} = \{\tilde{y}[1],\dots,\tilde{y}[r]\}. \tag{17}$$

In the vector form, Equation (16) is described by

$$\mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)}) = \mathbf{0}. \tag{18}$$

Equation (21) defines an implicit function $\lambda(\tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)})$. The existence and properties of this implicit function depend on the properties of the Jacobian matrix

$$J_\lambda(\lambda \mid \tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)}) = \left[\frac{\partial W_t}{\partial \lambda_i} \mid (t,i) = 1,\dots,r\right], \tag{19}$$

which has the elements

$$\frac{\partial W_t}{\partial \lambda_i} = \int_{Q} (\varphi(x[t],\theta) - \tilde{y}[t])\varphi(x[i],\theta) \sum_{j=1}^{r} \exp\left(-\sum_{j=1}^{r} \lambda_j \varphi(x[j],\theta)\right) d\theta. \tag{20}$$

Theorem 1. *Let the next conditions be valid (assume that):*

- *The function $\varphi(\mathbf{x}^{(r)},\theta)$ is continuous in all variables.*
- *For any $(\mathbf{x}^{(r)},\tilde{\mathbf{y}}^{(r)}) \in R^r \times R^r$,*

$$\det J_\lambda(\lambda \mid \tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)}) \ne 0, \tag{21}$$

$$\lim_{\|\lambda\| \to \infty} \mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)}) = \pm\infty. \tag{22}$$

Then, there exists a unique implicit function $\lambda(\tilde{\mathbf{y}}^{(r)},\mathbf{x}^{(r)},)$ defined on $R^r \times R^r$.

Proof of Theorem 1. Due to the first assumption, the continuous function $\mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ induces the vector field $\Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) = \mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ in the space $R^r \times R^r$.

We choose an arbitrary vector $\mathbf{u}$ in R^r and define the vector field

$$\Pi_{\mathbf{u}}(\lambda) = \Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) - \mathbf{u}.$$

By condition (22), the field $\Pi_{\mathbf{u}}(\lambda)$ with a fixed vector $\mathbf{u}$ has no zeros on the spheres $\|\lambda\| = \varrho$ of a sufficiently large radius ϱ.

Hence, a rotation is well defined on the spheres $\|\lambda\| = \varrho$ of a sufficiently large radius ϱ. For details, see [27].

Consider the two vector fields

$$\Pi_{\mathbf{u}^{(1)}}(\lambda) = \Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) - \mathbf{u}^{(1)}, \quad \Pi_{\mathbf{u}^{(2)}}(\lambda) = \Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) - \mathbf{u}^{(2)}.$$

These vector fields are homotopic on the spheres of a sufficiently large radius, i.e., the field

$$\Omega(\lambda) = \alpha \Pi_{\mathbf{u}^{(1)}}(\lambda) + (1-\alpha)\Pi_{\mathbf{u}^{(2)}}(\lambda) = \Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) - [\alpha \mathbf{u}^{(1)} + (1-\alpha)\mathbf{u}^{(2)}]$$

has no zeros on the spheres of a sufficiently large radius for any $\alpha \in [0,1]$. Homotopic fields have identical rotations [27]:

$$\gamma(\Pi_{\mathbf{u}^{(1)}}(\lambda)) = \gamma(\Pi_{\mathbf{u}^{(2)}}(\lambda)).$$

The vector fields $\Pi_{\mathbf{u}^{(1)}}(\lambda)$ and $\Pi_{\mathbf{u}^{(2)}}(\lambda)$ are nondegenerate on the spheres of a sufficiently large radius; in the ball $\|\lambda\| \le \varrho_1 < \varrho$, however, each of them may have a number of singular points. We denote by $\kappa(\mathbf{u}^{(1)})$ and $\kappa(\mathbf{u}^{(2)})$ the numbers of singular points of the vector fields $\Pi_{\mathbf{u}^{(1)}}(\lambda)$ and $\Pi_{\mathbf{u}^{(2)}}(\lambda)$, respectively. As the vector fields are homotopic,

$$\kappa(\mathbf{u}^{(1)}) = \kappa(\mathbf{u}^{(2)}) = \kappa.$$

In view of (21), these singular points are isolated.

Now, let us utilize the index of a singular point introduced in [27]:

$$\text{ind}\,(\lambda^0) = (-1)^{\beta(\lambda^0)},$$

where $\beta(\lambda^0)$ is the number of eigenvalues of the matrix $\Pi'_{\mathbf{u}}(\lambda^0) = J_\lambda(\lambda^0 \mid, \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ with the negative real part. By definition, the value of this index depends not on the magnitude of $\beta(\lambda^0)$, but on its parity. Due to condition (21), all singular points have the same parity. Really, $J_\lambda(\lambda^0 \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)}) \ne 0$, and hence, for any $\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)} \in R^r \times R^r$, the eigenvalues of the matrix $J_\lambda(\lambda^0 \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ may move from the left half-plane to the right one in pairs only: Real eigenvalues are transformed into pairs of complex–conjugate ones, passing through the imaginary axis.

In view of this fact, the rotation of the homotopic fields (20) is given by

$$\gamma(\Pi_{\mathbf{u}}) = \kappa(-1)^{\beta},$$

where β is the number of eigenvalues of the matrix $\Pi'_{\mathbf{u}}(\lambda)$ for some singular point.

It remains to demonstrate that the vector field $\Pi_{\mathbf{u}}(\lambda)$ has a unique singular point in the ball $\|\lambda\| \le \varrho_1 < \varrho$. Consider the equation

$$\Pi_{\mathbf{u}}(\lambda) = \Phi_{(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})}(\lambda) - \mathbf{u} = 0.$$

Assume that for each fixed pair $(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$, this equation has κ singular points, i.e., the functions $\lambda^{(1)}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)}), \ldots, \lambda^{(\kappa)}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$. Therefore, it defines a multivalued function $\lambda(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$, whose κ branches are isolated (the latter property follows from the isolation of

the singular points). Due to condition (21), each of the branches $\lambda^{(i)}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ defines an open set in the space R^r, and

$$\bigcup_{i=1}^{\kappa} \lambda^{(i)}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)}) = R^r.$$

This is possible if and only if $\kappa = 1$. Hence, for each pair $(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ from $R^r \times R^r$, there exists a unique function $\lambda^*(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ for which the function $\mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ will vanish. □

Theorem 2. *Under the assumptions of Theorem 1, the function $\lambda(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ is real analytical in all variables.*

Proof of Theorem 2. From (15), it follows that the function $\mathbf{W}(\lambda \mid \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ is analytical in all variables. Therefore, the left-hand side of Equation (15) can be expanded into the generalized Taylor series [26], and the solution can be constructed in the form of the generalized Taylor series as well. The power elements of this series are determined using a recursive procedure. □

5. Asymptotic Efficiency of RME Estimates

The RME estimate yields the entropy-optimal PDFs (10) for the arrays of input and output data, each of size r. For the sake of convenience, consider the PDFs parametrized by the exponential Lagrange multipliers $z = \exp(-\lambda)$. Then, equalities (10) take the form

$$\begin{aligned} P^*\left(\theta, \mathbf{z}(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\right) &= \frac{\prod_{j=1}^{r}[z_j(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})]^{\varphi(x[j],\theta)}}{\int\limits_{\Theta} \prod_{j=1}^{r}[z_j(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})]^{\varphi(x[j],\theta)} d\theta}, \\ Q_t^*(\xi[t], z_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})) &= \frac{[z_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})]^{\xi[t]}}{\int\limits_{\Xi_t} [z_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})]^{\xi[t]} d\xi[t]}, \quad t = 1, \ldots, r. \end{aligned} \tag{23}$$

Consequently, the structure of the PDF significantly depends on the values of the exponential Lagrange multipliers $\mathbf{z}$, which, in turn, depend on the data arrays $\mathbf{y}^{(r)}$ and $\mathbf{x}^{(r)}$.

Definition 1. *The estimates $P^*(\theta, \mathbf{z}^*)$ and $Q_t^*(\xi[t], z_t^*)$ are said to be asymptotically efficient if*

$$\begin{aligned} \lim_{r\to\infty} P^*\left(\theta, \mathbf{z}(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})\right) &= P^*(\theta, \mathbf{z}^*), \\ \lim_{r\to\infty} Q_t^*(\xi[t], z_t(\mathbf{y}^{(r)}, \mathbf{x}^{(r)})) &= Q_t^*(\xi[t], z_t^*), \quad t = 1, \ldots, r; \end{aligned} \tag{24}$$

where

$$\mathbf{z}^* = \lim_{r\to\infty} \mathbf{z}(\mathbf{y}^{(r)}, \mathbf{x}^{(r)}). \tag{25}$$

Consider the empirical balance Equation (21), written in terms of the exponential Lagrange multipliers:

$$\Phi_t(\mathbf{z}, \tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)}) = \int_{\Theta} \prod_{j=1}^{r} [z_j(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})]^{\varphi(x[j],\theta)} (\varphi(x[t], \theta) - \tilde{y}[t]) d\theta = 0, \quad t = 1, \ldots, r. \tag{26}$$

As has been demonstrated above, Equation (26) defines an implicit analytical function $\mathbf{z} = \mathbf{z}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ for $(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)}) \in R^r \times R^r$.

Differentiating the left- and right-hand sides of these equations with respect to $\tilde{\mathbf{y}}^{(r)}$ and $\mathbf{x}^{(r)}$ yields

$$\begin{aligned} \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{y}}^{(r)}} &= -\left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \frac{\partial \Phi}{\partial \tilde{\mathbf{y}}^{(r)}}, \\ \frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}} &= -\left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \frac{\partial \Phi}{\partial \mathbf{x}^{(r)}}. \end{aligned} \tag{27}$$

Then, passing to the norms and using the inequality for the norm of the product of matrices [28], we obtain the equalities

$$\begin{aligned} 0 \le \left\| \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{y}}^{(r)}} \right\| &\le \left\| \left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \right\| \left\| \frac{\partial \Phi}{\partial \tilde{\mathbf{y}}^{(r)}} \right\|, \\ 0 \le \left\| \frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}} \right\| &\le \left\| \left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \right\| \left\| \frac{\partial \Phi}{\partial \mathbf{x}^{(r)}} \right\|. \end{aligned} \tag{28}$$

Both of the inequalities incorporate the norm of the inverse matrix $\left\| \left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \right\|$.

Lemma 1. *Let a square matrix A be nonsingular, i.e.,* $\det A \neq 0$. *Then, there exists a constant $\alpha > 1$ such that*

$$\frac{1}{\|A\|} \le \|A^{-1}\| \le \frac{\alpha}{\|A\|}. \tag{29}$$

Proof of Lemma 1. Since the matrix A is nondegenerate, the elements $a_{ik}^{(-1)}$ of the inverse matrix A^{-1} can be expressed in terms of the algebraic complement (adjunct) of the element a_{ki} in the determinant of the matrix A [28]:

$$a_{ik}^{(-1)} = \frac{A_{ki}}{\det A}, \qquad (k,i) = 1, \dots, r,$$

and they are bounded:

$$a_{ik}^{(-1)} \le M < \infty, \qquad \|A^{-1}\| < \infty.$$

Hence, there exists a constant $\alpha > 1$ for which inequality (29) is satisfied. □

Lemma 1 can be applied to the norm $\left\| \left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \right\|$ of the inverse matrix. As a result,

$$\left(\left\| \frac{\partial \Phi}{\partial \mathbf{z}} \right\| \right)^{-1} \le \left\| \left[\frac{\partial \Phi}{\partial \mathbf{z}}\right]^{-1} \right\| \le \alpha \left(\left\| \frac{\partial \Phi}{\partial \mathbf{z}} \right\| \right)^{-1}, \tag{30}$$

where

$$\left\| \frac{\partial \Phi}{\partial \mathbf{z}} \right\| = r \max_{t,j} \left| \frac{\partial \Phi_t}{\partial z_j} \right|. \tag{31}$$

Lemma 2. *Let*

$$\| \frac{\partial \Phi}{\partial \tilde{\mathbf{y}}^{(r)}} \| \le \varrho < \infty, \qquad \| \frac{\partial \Phi}{\partial \mathbf{x}^{(r)}} \| \le \omega < \infty. \tag{32}$$

Then,

$$\lim_{r \to \infty} \| \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{y}}^{(r)}} \| = \lim_{r \to \infty} \| \frac{\partial \mathbf{z}}{\partial \mathbf{x}^{(r)}} \| = 0. \tag{33}$$

Proof of Lemma 2. According to (28), (31), and (32) we have:

$$\|\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{y}}^{(r)}}\| \le \frac{1}{r}\Big(\frac{\varrho}{b}\Big), \quad \|\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{x}}^{(r)}}\| \le \frac{1}{r}\Big(\frac{\omega}{b}\Big),$$

where $b = \max_{t,j}\Big|\frac{\partial \Phi_t}{\partial z_j}\Big|$.

Whence, it follows that for the sample length $r \to \infty$, the norms of relevant Jacobians tend to zero, and function $\mathbf{z} = \mathbf{z}(\tilde{\mathbf{y}}^{(r)}, \mathbf{x}^{(r)})$ tends to the vector $\mathbf{z}^*$ (25). □

6. Thermokarst Lake Area Evolution in Western Siberia: RME Estimation and Testing

Permafrost zones, which occupy a significant part of the Earth's surface, are the locales of thermokarst lakes, which accumulate greenhouse gases (methane and carbon dioxide). These gases make a considerable contribution to global climate change.

The source data in studies of the evolution of thermokarst lake areas are acquired through remote sensing of the Earth's surface and ground measurements of meteorological parameters [29,30].

The state of thermokarst lakes is characterized by their total area $S[t]$ in a given region, measured in hectares (ha), and the factors influencing thermokarst formations—the average annual temperatures $T[t]$, measured in Celsius (C°), and the annual precipitation $R[t]$, measured in millimeters (mm), where t denotes the calendar year.

We used the remote sensing data and ground measurements of the meteorological parameters for a region of Western Siberia between 65° N–70° N and 65° E–95° E that were presented in [31]. We divided the available time series into two groups, which formed the training collection $\mathcal{L}$ ($t = 0, \dots, 24$) and the testing collection $\mathcal{T}$ ($t = 25, \dots, 35$).

6.1. RME Estimation of Model Parameters and Measurement Noises

The temporal evolution of the lake area $S[t]$ is described by the following dynamic regression equation with two influencing factors, the average annual temperature $T[t]$ and the annual precipitation $R[t]$:

$$\begin{aligned} \hat{S}[t] &= a_0 + \sum_{k=1}^{p} a_k \hat{S}[t-k] + a_{(p+1)} T[t] + a_{(p+2)}(R[t], \\ \hat{v}[t] &= \hat{S}[t] + \xi[t]. \end{aligned} \tag{34}$$

The model parameters and measurement noises are assumed to be random and of the interval type:

$$\begin{aligned} &a_k \in \mathcal{A}_k = [a_k^-, a_k^+], \quad k = 0, dots, p+2, \\ &\mathbf{a} = \{a_0, \dots, a_p, a_{p+1}, a_{p+2}\} \in \mathcal{A} = \bigcup_{k=0}^{p+2} \mathcal{A}_k. \end{aligned}$$

The probabilistic properties of the parameters are characterized by a PDF $P(\mathbf{a})$.

The variable $\hat{v}[t]$ is the observed output of the model, and the values of the random measurement noise $\xi[t]$ at different time instants t may belong to different ranges:

$$\xi[t] \in \Xi_t = [\xi^-[t], \xi^+[t]], \tag{35}$$

with a PDF $Q_t(\xi[t])$, $(t = 0, \dots, N)$, where N denotes the length of the observation interval. The order $p = 4$ and the parameter ranges for the dynamic randomized regression model (34) (see Table 1 below) were calculated based on real data using the empirical correlation functions and the least-square estimates of the residual variances.

Table 1. Parameter ranges for the model.

$\mathbf{a}$	a_0	a_1	a_2	a_3	a_4	a_5	a_6
$\mathbf{a}^-$	−0.50	−0.14	−0.49	−0.53	−0.44	0.46	0.19
$\mathbf{a}^+$	0.07	0.52	0.20	0.19	0.19	1.14	0.88

For the training collection $\mathcal{L}$, the model can be written in the vector–matrix form

$$\begin{aligned} \hat{\mathbf{S}} &= \hat{\mathbb{S}}\mathbf{a} + a_5\mathbf{T} + a_6\mathbf{R}, \\ \hat{\mathbf{v}} &= \hat{\mathbf{S}} + \xi, \end{aligned} \tag{36}$$

with the matrix

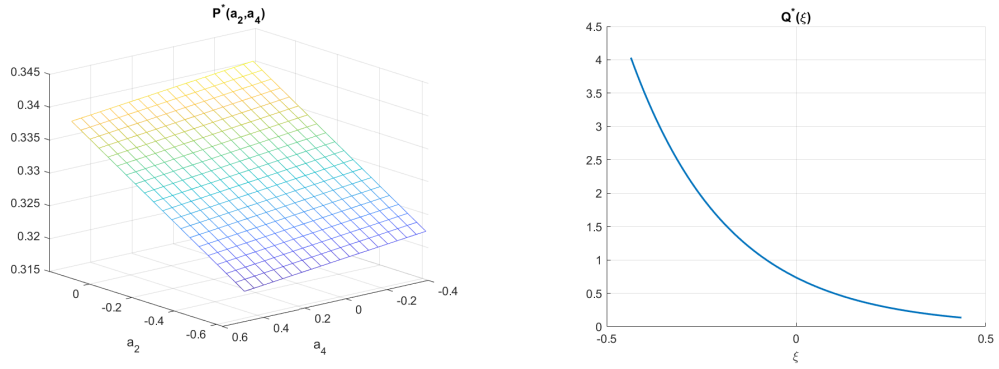
$$\hat{\mathbb{S}} = \begin{pmatrix} 1 & \hat{S}[3] & \cdots & \hat{S}[0] \\ 1 & \hat{S}[4] & \cdots & \hat{S}[1] \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \hat{S}[23] & \cdots & \hat{S}[20] \end{pmatrix} \tag{37}$$

and the vectors $\hat{\mathbf{S}} = [\hat{S}[4], \ldots, \hat{S}[24]], \mathbf{T} = [T[4], \ldots, T[24]], \mathbf{R} = [R[4], \ldots, R[24]]$, and $\hat{\mathbf{v}} = [v[4], \ldots, v[24]]$; $\xi = [\xi[4], \ldots, \xi[24]]$.

The RME estimation procedure yielded the following entropy-optimal PDFs of the model parameters (36) and measurement noises:

$$\begin{aligned} &P^*(\mathbf{a}, \lambda) = \prod_{k=0}^{6} \frac{\exp(-q_k a_k)}{\mathcal{P}_k(\lambda)}, \ \mathcal{P}_k(\lambda) = \int_{\mathcal{A}_\parallel} \exp(-q_k a_k) da_k, \\ &q_0 = \sum_{t=4}^{24} \lambda_n, \quad q_k = \sum_{t=p}^{24} \lambda_n \, S[t-k], \quad k = 1, \ldots, 4, \\ &q_5 = \sum_{t=4}^{24} \lambda_t T[t], \quad q_6 = \sum_{t=p}^{24} \lambda_t R[t], \\ &Q^*(\xi, \bar{\lambda}) = \frac{\exp(-\bar{\lambda}\,\xi)}{\mathcal{Q}}, \quad \mathcal{Q} = \int_{\Xi} \exp(-\bar{\lambda}\,\xi) d\xi, \quad \bar{\lambda} = \frac{q_0}{20}. \end{aligned} \tag{38}$$

Note that $S[t-k]$, $T[t]$, and $R[t]$ are the data from the collection $\mathcal{L}$. The two-dimensional sections of the function $P^*(\mathbf{a})$ and the function $Q^*(\xi)$ are shown in Figure 1.



(**a**) Two-dimensional section of function $P^*(\mathbf{a})$ (**b**) Function $Q^*(\xi)$.

Figure 1. Two-dimensional section of the function P^* and the function Q^*.

6.2. Testing

Testing was performed using the data from the collection $\mathcal{T}$, which included the lake area $S[t]$, the average annual temperature $T[t]$, and the annual precipitation $R[t]$, $t = 25, \ldots, 35$. An ensemble of trajectories of the model's observed output $v[t]$ was generated using Monte Carlo simulations and sampling of the entropy-optimal PDFs

$P^*(\mathbf{a})$, $Q^*\xi$ on the testing interval. In addition, the trajectory of the empirical means $\bar{v}[t]$ and the dimensions of the empirical standard deviation area were calculated.

The quality of RME estimation was characterized by the absolute and relative errors:

$$AbsErr = \sqrt{\sum_{t=26}^{35} \left(S[t] - \bar{v}[t]\right)^2} = 0.3446, \tag{39}$$

$$RelErr = \frac{\sqrt{\sum_{t=26}^{35} \left(S[t] - \bar{v}[t]\right)^2}}{\sqrt{\sum_{t=26}^{35} S^2[t]} + \sqrt{\sum_{t=26}^{35} \bar{v}^2[t]}} = 0.0089. \tag{40}$$

The generated ensemble of the trajectories is shown in Figure 2.

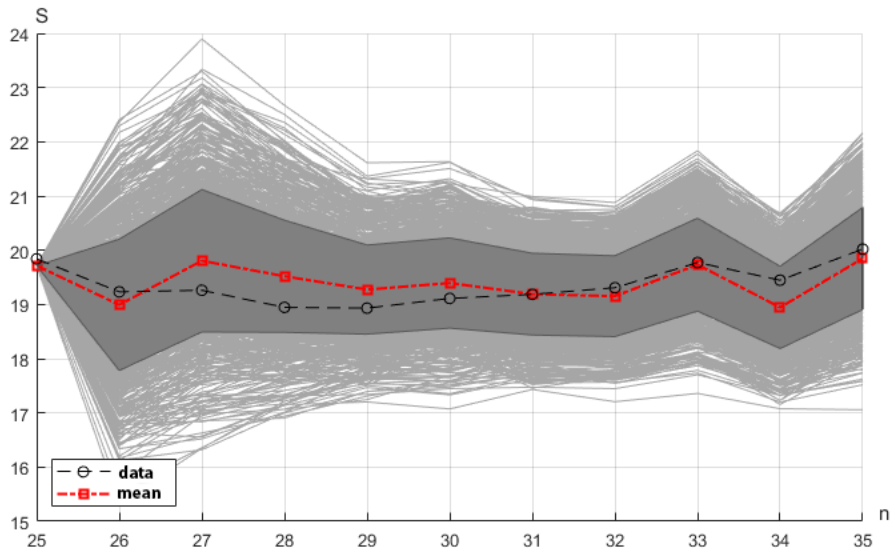


Figure 2. Ensemble of the trajectories (gray domain), the standard deviation area (dark gray domain), the empirical mean trajectory, and the lake area data.

7. Discussion

Given an available data collection, the RME procedure allows estimation of the PDFs of a model's random parameters under measurement noises corresponding to the maximum uncertainty (maximum entropy). In addition, this procedure needs no assumptions about the structure of the estimated PDFs or the statistical properties of the data and measurement noises.

An entropy-optimal model can be simulated by sampling the PDFs to generate an empirical ensemble of a model's output trajectories and to calculate its empirical characteristics (the mean and median trajectories, the standard deviation area, interquartile sets, and others).

The RME procedure was illustrated with an example of the estimation of the parameters of a linear regression model for the evolution of the thermokarst lake area in Western Siberia. In this example, the procedure demonstrated a good estimation accuracy.

However, these positive features of the procedure were achieved with computational costs. Despite their analytical structure, the RME estimates of the PDFs depend on Lagrange multipliers, which are determined by solving the balance equations with the so-called integral components (the mathematical expectations of random parameters and measurement noises). Calculating the values of multidimensional integrals may require appropriate computing resources.

8. Conclusions

The problem of randomized maximum entropy estimation of a probability density function based on real available data has been formulated and solved. The developed estimation algorithm (the RME algorithm) finds the conditional maximum of an information entropy functional on a set of admissible probability density functions characterized by the empirical balance equations for Lagrange multipliers. These equations define an implicit dependence of the Lagrange multipliers on the data collection. The existence of such an implicit function for any values in a data collection has been established. The function's behavior for a data collection of a greater size has been studied, and the asymptotic efficiency of the RME estimates has been proved.

The positive features of RME estimates have been illustrated with an example of estimation and testing a linear dynamic regression model of the evolution of the thermokarst lake area in Western Siberia with real data.

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation, project no. 075-15-2020-799.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
2. Witten, I.H.; Eibe, F. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Heidelberg, Germany, 2005.
3. Bishop, C.M. *Pattern Recognition and Machine Learning. Series: Information Theory and Statistics*; Springer: New York, NY, USA, 2006.
4. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.
5. Vorontsov, K.V. *Mathematical Methods of Learning by Precedents: A Course of Lectures*; Moscow Institute of Physics and Technology: Moscow, Russia, 2013.
6. Goldberger, A.S. *A Course in Econometrics*; Harvard University Press: Cambridge, UK, 1991.
7. Aivazyan, S.A.; Enyukov, I.S.; Meshalkin, L.D. *Prikladnaya Statistika: Issledovanie Zavisimostei (Applied Statistics: Study of Dependencies)*; Finansy i Statistika: Moscow, Russia, 1985.
8. Lagutin, M.B. *Naglyadnaya Matematicheskaya Statistika (Visual Mathematical Statistics)*; BINOM, Laboratoriya Znanii: Moscow, Russia, 2013.
9. Roussas, G. *A Course of the Mathematical Statistics*; Academic Press: San Diego, CA, USA, 2015.
10. Malouf, R. A comparison of algorithms for maximum entropy parameters estimation. In Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002), Taipei, Taiwan, 31 August–1 September 2002; Volume 20, pp. 1–7.
11. Borwein, J.; Choksi, R.; Marechal, P. Probability distribution of assets inferred from option prices via principle of maximum entropy. *SIAM J. Optim.* **2003**, *14*, 464–478. [CrossRef]
12. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1997.
13. Golan, A. Information and Entropy econometrics—A review and synthesis. *Found. Trends Econom.* **2008**, *2*, 1–145. [CrossRef]
14. Csiszar, I.; Matus, F. On minimization of entropy functionals under moment constraints. In Proceedings of the IEEE International Symposium on Information Theory, Toronto, ON, Canada, 6–11 July 2008.
15. Loubes, J.-M. Approximate maximum entropy on the mean for instrumental variable regression. *Stat. Probab. Lett.* **2012**, *82*, 972–978. [CrossRef]
16. Borwein, J.M.; Lewis, A.S. Partially-finite programming in L_1 and existence of maximum entropy estimates. *SIAM J. Optim.* **1993**, *3*, 248–267. [CrossRef]
17. Burg, J.P. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics* **1972**, *37*, 375–376. [CrossRef]
18. Christakos, G. A Bayesian/maximum entropy view to the spatial estimation problem. *Math. Geol.* **1990**, *22*, 763–777. [CrossRef]
19. Singh, V.P.; Guo H. Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy. *Hydrol. Sci. J.* **1994**, *40*, 165–181. [CrossRef]
20. Popkov, Y.S.; Dubnov, Y.A.; Popkov, A.Y. Randomized machine learning: Statement, solution, applications. In Proceedings of the 2016 IEEE 8th International Conference on Intelligent Systems (IS), Sofia, Bulgaria, 4–6 September 2016. [CrossRef]
21. Popkov, A.Y.; Popkov, Y.S. New methods of entropy-robust estimation for randomized models under limited data. *Entropy* **2014**, *16*, 675–698. [CrossRef]
22. Krasnosel'skii, M.A.; Vainikko, G.M.; Zabreyko, R.P.; Ruticki, Y.B.; Stet'senko, V.V. *Approximate Solutions of Operator Equations*; Wolters-Noordhoff Publishing: Groningen, The Netherlands, 1972. [CrossRef]

23. Ioffe, A.D.; Tikhomirov, V.M. *Theory of Extremal Problems*; Elsevier: New York, NY, USA, 1974.
24. Alekseev, V.M.; Tikhomirov, V.M.; Fomin, S.V. *Optimal Control*; Springer: Boston, MA, USA, 1987.
25. Kaashoek, M.A.; van der Mee, C. *Recent Advances in Operator Theory and Its Applications*; Birkhäuser Basel: Basel, Switzerland, 2006.
26. Kolmogorov, A.N.; Fomin, S.V. *Elements of the Theory of Functions and Functional Analysis*; Dover Publication: New York, NY, USA, 1999.
27. Krasnoselskii, M.A.; Zabreiko, P.P. *Geometrical Methods of Nonlinear Analysis*; Springer: Berlin, Germany; New York, NY, USA, 1984.
28. Gantmacher, F.R.; Brenner, J.L. *Applications of the Theory of Matrices*; Dover: New York, NY, USA, 2005.
29. Riordan, B.; Verbula, D.; McGruire, A.D. Shrinking ponds in subarctic Alaska based on 1950–2002 remotely sensed images. *J. Geophys. Res.* **2006**, *111*, G04002. [CrossRef]
30. Kirpotin, S.; Polishchuk, Y.; Bruksina, N. Abrupt changes of thermokarst lakes in Western Siberia: Impacts of climatic warming on permafrost melting. *Int. J. Environ. Stud.* **2009**, *66*, 423–431. [CrossRef]
31. Western Siberia Thermokarsk Lakes Dataset. Available online: https://cloud.uriit.ru/index.php/s/0DOrxL9RmGqXsV0 (accessed on 20 February 2021).