Journal of Machine Learning Research 16 (2015) 1683-1730 Submitted 7/15; Published 9/15

In memory of Alexey Chervonenkis

V-Matrix Method of Solving Statistical Inference Problems

Vladimir Vapnik VLADIMIR.VAPNIK@GMAIL.COM
Columbia University
New York, NY 10027, USA
Facebook AI Research
New York, NY 10017, USA

Rauf Izmailov RIZMAILOV@APPCOMSCI.COM
Applied Communication Sciences
Basking Ridge, NJ 07920-2021, USA

Editors: Alex Gammerman and Vladimir Vovk

Abstract

This paper presents direct settings and rigorous solutions of the main Statistical Inference problems. It shows that rigorous solutions require solving multidimensional Fredholm integral equations of the first kind in the situation where not only the right-hand side of the equation is an approximation, but the operator in the equation is also defined approximately. Using Stefanuyk-Vapnik theory for solving such ill-posed operator equations, constructive methods of empirical inference are introduced. These methods are based on a new concept called V-matrix. This matrix captures geometric properties of the observation data that are ignored by classical statistical methods.

Keywords: conditional probability, regression, density ratio, ill-posed problem, mutual information, reproducing kernel Hilbert space · function estimation, interpolation function, support vector machines, data adaptation, data balancing, conditional density

1. Basic Concepts of Classical Statistics

In the next several sections, we describe main concepts of Statistics. We first outline these concepts for the one-dimensional case and then generalize them for the multidimensional case.

1.1 Cumulative Distribution Function

The basic concept of *Theoretical Statistics* and *Probability Theory* is the so-called *Cumulative Distribution Function* (CDF)

$$F(x) = P\{X \leq x\}.$$

This function defines the probability of the random variable X not exceeding x. Different CDFs describe different statistical environments, so CDF (defining the probability measure) is the main characteristic of the random events. In this paper, we consider the important case when $F(x)$ is a *continuous* function.

1.2 General Problems of Probability Theory and Statistics

The general problem of Probability Theory can be defined as follows:

> *Given a cumulative distribution function $F(x)$, describe outcomes of random experiments for a given theoretical model.*

The general problem of Statistics can be defined as follows:

> *Given iid observations of outcomes of the same random experiments, estimate the statistical model that defines these observations.*

In Section 2, we discuss several main problems of Statistics. Next, we consider the basic one: estimation of CDF.

1.3 Empirical Cumulative Distribution Functions

In order to estimate CDF, one introduces the so-called *Empirical Cumulative Distribution function* (ECDF) constructed for iid observations obtained according to $F(x)$:

$$X_1, ..., X_\ell.$$

The ECDF function has the form

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x - X_i),$$

where $\theta(x - X_i)$ is the step-function

$$\theta(x - X_i) = \begin{cases} 1, & \text{if } x \geq X_i, \\ 0, & \text{if } x < X_i. \end{cases}$$

Classical statistical theory is based on convergence of ECDF converges to CDF when the number ℓ of observations increases.

1.4 The Glivenko-Cantelli Theorem and Kolmogorov Type Bounds

In 1933, the following theorem was proven (Glivenko-Cantelli theorem).

Theorem. *Empirical cumulative distribution functions converge uniformly to the true cumulative distribution function:*

$$\lim_{\ell\to\infty} P\{\sup_x |F(x) - F_\ell(x)| \geq \varepsilon\} = 0, \quad \forall\varepsilon > 0.$$

In 1933, Kolmogorov derived asymptotical exact rate of convergence of ECDF to CDF for continuous functions $F(x)$:

$$\lim_{\ell\to\infty} P\{\sqrt{\ell}\sup_x |F(x) - F_\ell(x)| \geq \varepsilon\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\{-2\varepsilon^2 k^2\}. \tag{1}$$

Later, Dvoretzky, Kiefer, Wolfowitz, and Massart showed the existence of exponential type of bounds for any ℓ:

$$P\{\sup_x |F(x) - F_\ell(x)| \geq \varepsilon\} \leq 2\exp\{-2\varepsilon^2\ell\}. \tag{2}$$

Bound (2) is defined by the first term of the right-hand side of Kolmogorov asymptotic equality (1).

Glivenko-Cantelli theorem and bounds (1), (2) can be considered as a foundation of statistical science since they claim that:

1. It is possible to estimate the true statistical distribution from iid data.
2. The ECDF strongly converges to the true CDF, and this convergence is fast.

1.5 Generalization to Multidimensional Case

Let us generalize the main concepts described above to the multidimensional case. We start with CDF.

Joint cumulative distribution function. For the multivariate random variable $x = (x^1, ..., x^d)$, the joint cumulative distribution function $F(x)$, $x \in R^d$ is defined by the function

$$F(x) = P\{X^1 \leq x^1, ..., X^d \leq x^d\}. \tag{3}$$

As in the one-dimensional case, the main problem of Statistics is as follows: estimate CDF, as defined in (3), based on random multivariate iid observations

$$X_1, ..., X_\ell, \quad X_i \in R^d, \ \ i = 1, \dots, \ell..$$

In order to solve this problem, one uses the same idea of empirical distribution function

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - X_i),$$

where $x = (x^1, ..., x^d) \in R^d$, $\ X_i = (X_i^1, ..., X_i^d) \in R^d$ and

$$\theta(x - X_i) = \prod_{k=1}^{d}\theta(x^k - X_i^k).$$

Note that

$$F(x) = E_u\theta(x - u) = \int \theta(x - u)dF(u),$$

and the generalized (for the multidimensional case) Glivenko-Cantelli theorem has the form

$$\lim_{\ell\to\infty} P\left\{\sup_x \left|E_u\theta(x-u) - \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - X_i)\right| \geq \varepsilon\right\} = 0.$$

This equation describes the uniform convergence of the empirical risks to their expectation over vectors $u \in R^d$ for the parametric set of multidimensional step functions $\theta(x-u)$ (here

$x, u \in R^d$, and x is a vector of parameters). Since VC dimension of this set of functions is equal[1] to one, according to the VC theory (Vapnik and Chervonenkis, 1974), (Vapnik, 1995), (Vapnik, 1998), the corresponding rate of convergence is bounded as follows:

$$P\left\{\sup_x \left|E_u\theta(x-u) - \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x-X_i)\right| \geq \varepsilon\right\} \leq \exp\left\{-\left(\varepsilon^2 - \frac{\ln \ell}{\ell}\right)\ell\right\}. \qquad (4)$$

According to this bound, for sufficiently large values of ℓ, the convergence of ECDF to the actual CDF does not depend on the dimensionality of the space. This fact has important consequences for Applied Statistics.

2. Main Problems of Statistical Inference

The main target of statistical inference theory is estimation (from the data) of specific models of random events, namely:

1. conditional probability function;
2. conditional density function;
3. regression function;
4. density ratio function.

2.1 Conditional Density, Conditional Probability, Regression, and Density Ratio Functions

Let $F(x)$ be a cumulative distribution function of random variable x. We call non-negative function $p(x)$ the probability density function if

$$\int_{-\infty}^{x} p(x^*)dx^* = F(x).$$

Similarly, let $F(x,y)$ be the joint probability distribution function of variables x and y. We call non-negative $p(x,y)$ the joint probability density function of two variables x and y if

$$\int_{-\infty}^{y}\int_{-\infty}^{x} p(x^*, y^*)dx^*dy^* = F(x,y).$$

1. Let $p(x,y)$ and $p(x)$ be probability density functions for pairs (x,y) and vectors x. Suppose that $p(x) > 0$. The function

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

is called the *Conditional Density Function.* It defines, for any fixed $x = x_0$, the probability density function $p(y|x = x_0)$ of random value $y \in R^1$. The estimation of the conditional density function from data

$$(y_1, X_1), ..., (y_\ell, X_\ell) \qquad (5)$$

[1] Since the set of d-dimensional parametric (with respect to parameter x) functions $\theta(x-u)$ can shatter, at most, one vector.

is the most difficult problem in our list of statistical inference problems.

2. Along with estimation of the conditional density function, the important problem is to estimate the so-called *Conditional Probability Function.* Let variable y be discrete, say, $y \in \{0, 1\}$. The function defined by the ratio

$$p(y = 1|x) = \frac{p(x, y = 1)}{p(x)}, \quad p(x) > 0$$

is called *Conditional Probability Function.* For any given vector $x = x_0$, this function defines the probability that y is equal to one; correspondingly, $p(y = 0|x = x_0) = 1 - p(y = 1|x = x_0)$. The problem is to estimate the conditional probability function, given data (5) where $y \in \{0, 1\}$.

3. As mentioned above, estimation of the conditional density function is a difficult problem; a much easier problem is the problem of estimating the so-called *Regression Function* (conditional expectation of the variable y):

$$r(x) = \int y p(y|x) dy,$$

which defines expected value $y \in R^1$ for a given vector x.

4. In this paper, we also consider a problem, which is important for applications: estimating the ratio of two probability densities (Sugiyama et al., 2012). Let $p_{\text{num}}(x)$ and $p_{\text{den}}(x) > 0$ be two different density functions (subscripts *num* and *den* correspond to numerator and denominator of the density ratio). Our goal is to estimate the function

$$R(x) = \frac{p_{\text{num}}(x)}{p_{\text{den}}(x)}$$

given iid data

$$X_1, ..., X_{\ell_{\text{den}}},$$

distributed according to $p_{\text{den}}(x)$, and iid data

$$X'_1, ..., X'_{\ell_{\text{num}}},$$

distributed according to $p_{\text{num}}(x)$.

In the next sections, we introduce direct settings for these four statistical inference problems.

2.2 Direct Constructive Setting for Conditional Density Estimation

By definition, conditional density $p(y|x)$ is the ratio of two densities

$$p(y|x) = \frac{p(x, y)}{p(x)}, \quad p(x) > 0 \tag{6}$$

or, equivalently,

$$p(y|x)p(x) = p(x, y).$$

This expression leads to the following equivalent one:

$$\int\int \theta(y-y')\theta(x-x')f(x',y')dF(x')dy' = F(x,y), \tag{7}$$

where $f(x,y) = p(y|x)$, function $F(x)$ is the cumulative distribution function of x and $F(x,y)$ is the joint cumulative distribution function of x and y.

Therefore, our setting of the condition density estimation problem is as follows:

> *Find the solution of integral equation (7) in the set of nonnegative functions $f(x,y) = p(y|x)$ when the cumulative probability distribution functions $F(x,y)$ and $F(x)$ are unknown but iid data*
>
> $$(y_1, X_1), ..., (y_\ell, X_\ell)$$
>
> *are given.*

In order to solve this problem, we use empirical estimates

$$F_\ell(x,y) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(y-y_i)\theta(x-X_i), \tag{8}$$

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x-X_i) \tag{9}$$

of the unknown cumulative distribution functions $F(x,y)$ and $F(x)$. Therefore, we have to solve an integral equation where not only its right-hand side is defined approximately ($F_\ell(x,y)$ instead of $F(x,y)$), but also the data-based approximation

$$A_\ell f(x,y) = \int\int \theta(y-y')\theta(x-x')f(x',y')dy'dF_\ell(x')$$

is used instead of the exact integral operator

$$Af(x,y) = \int\int \theta(y-y')\theta(x-x')f(x',y')dy'dF(u').$$

Taking into account (9), our goal is thus to find the solution of approximately defined equation

$$\sum_{i=1}^{\ell} \theta(x-X_i)\int_{-\infty}^{y} f(X_i,y')dy' \approx \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(y-y_i)\theta(x-X_i). \tag{10}$$

Taking into account definition (6), we have

$$\int_{-\infty}^{\infty} p(y|x)dy = 1, \quad \forall x \in \mathcal{X}.$$

Therefore, the solution of equation (10) has to satisfy the constraint $f(x,y) \geq 0$ and the constraint

$$\int_{-\infty}^{\infty} f(y',x)dy' = 1, \quad \forall x \in \mathcal{X}.$$

We call this setting *the direct constructive setting* since it is based on direct definition of conditional density function (7) and uses theoretically justified approximations (8), (9) of unknown functions.

2.3 Direct Constructive Setting for Conditional Probability Estimation

The problem of estimation of the conditional probability function can be considered analogously to the conditional density estimation problem. The conditional probability is defined as

$$p(y=1|x) = \frac{p(x,y=1)}{p(x)}, \quad p(x) > 0 \tag{11}$$

or, equivalently,

$$p(y=1|x)p(x) = p(x,y=1).$$

We can rewrite it as

$$\int \theta(x-x')f(x')dF(x') = F(x,y=1), \tag{12}$$

where $f(x) = p(y=1|x)$ and $F(x,y=1) = P\{X \leq x, y=1\}$.

Therefore, the problem of estimating the conditional probability is formulated as follows.

In the set of bounded functions $0 \leq f(x) \leq 1$, find the solution of equation (12) if cumulative distribution functions $F(x)$ and $F(x,y=1)$ are unknown but iid data

$$(y_1,X_1),...,(y_\ell,X_\ell), \qquad y \in \{0,1\}, \quad x \in \mathcal{X},$$

generated according to $F(x,y)$, are given.

As before, instead of unknown cumulative distribution functions we use their empirical approximations

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x-X_i), \tag{13}$$

$$F_\ell(x,y=1) = p_\ell F_\ell(x|y=1) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\theta(x-X_i), \tag{14}$$

where p_ℓ is the ratio of the number of examples with $y=1$ to the total number ℓ of the observations.

Therefore, one has to solve integral equation (12) with approximately defined right-hand side (13) and approximately defined operator (14):

$$A_\ell f(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x-X_i)f(X_i).$$

Since the probability takes values between 0 and 1, our solution has to satisfy the bounds

$$0 \leq f(x) \leq 1, \quad \forall x \in \mathcal{X}.$$

Also, definition (11) implies that

$$\int f(x)dF(x) = p(y=1),$$

where $p(y=1)$ is the probability of $y=1$.

2.4 Direct Constructive Setting for Regression Estimation

By definition, regression is the conditional mathematical expectation

$$r(x) = \int y p(y|x) dy = \int y \frac{p(x,y)}{p(x)} dy.$$

This can be rewritten in the form

$$r(x)p(x) = \int y p(x,y) dy. \tag{15}$$

From (15), one obtains the equivalent equation

$$\int \theta(x-x') r(x') dF(x') = \int \theta(x-x') \int y dF(x',y'). \tag{16}$$

Therefore, the direct constructive setting of regression estimation problem is as follows:

In a given set of functions $r(x)$, find the solution of integral equation (16) if cumulative probability distribution functions $F(x,y)$ and $F(x)$ are unknown but iid data (5) are given.

As before, instead of these functions, we use their empirical estimates. That is, we construct the approximation

$$A_\ell r(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x-X_i) r(X_i)$$

instead of the actual operator in (16), and the approximation of the right-hand side

$$F_\ell(x) = \frac{1}{\ell} \sum_{j=1}^{\ell} y_j \theta(x-X_j)$$

instead of the actual right-hand side in (16), based on the observation data

$$(y_1,X_1),...,(y_\ell,X_\ell), \qquad y \in R^1, \quad x \in \mathcal{X}. \tag{17}$$

2.5 Direct Constructive Setting of Density Ratio Estimation Problem

Let $F_{\text{num}}(x)$ and $F_{\text{den}}(x)$ be two different cumulative distribution functions defined on $\mathcal{X} \subset R^d$ and let $p_{\text{num}}(x)$ and $p_{\text{den}}(x)$ be the corresponding density functions. Suppose that $p_{\text{den}}(x) > 0, x \in \mathcal{X}$. Consider the ratio of two densities:

$$R(x) = \frac{p_{\text{num}}(x)}{p_{\text{den}}(x)}.$$

The problem is to estimate the ratio $R(x)$ when densities are unknown, but iid data

$$X_1,...,X_{\ell_{\text{den}}} \sim F_{\text{den}}(x), \tag{18}$$

generated according to $F_{\text{den}}(x)$, and iid data

$$X'_1,...,X'_{\ell_{\text{num}}} \sim F_{\text{num}}(x), \tag{19}$$

generated according to $F_{\text{num}}(x)$, are given.

As before, we introduce the constructive setting of this problem: solve the integral equation

$$\int \theta(x-u)R(u)dF_{\text{den}}(u) = F_{\text{num}}(x)$$

when cumulative distribution functions $F_{\text{den}}(x)$ and $F_{\text{num}}(x)$ are unknown, but data (18) and (19) are given. As before, we approximate the unknown cumulative distribution functions $F_{\text{num}}(x)$ and $F_{\text{den}}(x)$ using empirical distribution functions

$$F_{\ell_{\text{num}}}(x) = \frac{1}{\ell_{\text{num}}} \sum_{j=1}^{\ell_{\text{num}}} \theta(x - X'_j)$$

for $F_{\text{num}}(x)$, and

$$F_{\ell_{\text{den}}}(x) = \frac{1}{\ell_{\text{den}}} \sum_{j=1}^{\ell_{\text{den}}} \theta(x - X_j)$$

for $F_{\text{den}}(x)$.

Since $R(x) \geq 0$ and $\lim_{x\to\infty} F_{\text{num}}(x) = 1$, our solution has to satisfy the constraints

$$R(x) \geq 0, \quad \forall x \in \mathcal{X},$$

$$\int R(x)dF_{\text{den}}(x) = 1.$$

Therefore, all main empirical inference problems can be represented via (multidimensional) Fredholm integral equation of the first kind with approximately defined elements. Although approximations converge to the true functions, these problems are computationally difficult due to their ill-posed nature. Thus they require rigorous solutions.[2]

In Section 5, we consider methods for solving ill-posed operator equations, which we apply in Section 6 to our problems of inference. Before that, however, we present a general form for all statistical inference problems in the next subsections.

2.6 General Form of Statistical Inference Problems

Consider the multidimensional Fredholm integral equation

$$\int \theta(z - z')f(z')dF_A(z') = F_B(z),$$

where the kernel of operator equation is defined by the step function $\theta(z-z')$, the cumulative distribution functions $F_A(z)$ and $F_B(z)$ are unknown but the corresponding iid data

$$Z_1, ..., Z_{\ell_A} \sim F_A(z)$$

$$Z_1, ..., Z_{\ell_B} \sim F_B(z)$$

are given. In the different inference problems, the elements $f(z), F_A(z), F_B(z)$ of the equation have different meanings (Table 1):

[2] Various statistical methods exist for solving these inference problems. Our goal is to find general rigorous solutions that take into account all the available characteristics of the problems.

	Conditional density	**Conditional probability**	**Density ratio**	**Regression**
z	(x, y)	x	x	(x, y), where $y \geq 0$
$f(z)$	$p(y\|x)$	$p(y = 1\|x)$	$\dfrac{p_{num}(x)}{p_{den}(x)}$	$\hat{y}^{-1}R(x),\ (R(x) = \displaystyle\int yp(y\|x)dy)$
$F_A(z)$	$F(x)$	$F(x)$	$F_{num}(x)$	$F(x)$
$F_B(z)$	$F(x, y)$	$F(x\|y = 1)p(y = 1)$	$F_{den}(x)$	$\hat{y}^{-1}\displaystyle\int \theta(x - x')y'dF(x', y')$

Table 1: Vector z, solution $f(z)$, and functions $F_A(z)$, $F_B(z)$ for different statistical inference problems.

1. In the problem of conditional density estimation, vector z is the pair (x, y), the solution $f(z)$ is $p(y|x)$, the cumulative distribution function $F_A(z)$ is $F(x)$ and the cumulative distribution function $F_B(z)$ is $F(x, y)$.

2. In the problem of conditional probability $p(y = 1|x)$ estimation, vector z is x, the solution $f(z)$ is $p(y = 1|x)$, the cumulative distribution function $F_A(z)$ is $F(x)$, the cumulative distribution function $F_B(z)$ is $F(x|y = 1)p(y = 1)$, where $p(y = 1)$ is the probability of class $y = 1$.

3. In the problem of density ratio estimation, the vector z is x, the solution $f(z)$ is $p_{num}(x)/p_{den}(x)$, the cumulative function $F_A(z)$ is $F_{num}(x)$, the cumulative function $F_B(z)$ is $F_{den}(x)$.

4. In the problem of regression $R(x) = \int yp(y|x)dy$ estimation, the vector z is (x, y), where $y \geq 0$, the solution $f(z)$ is $\hat{y}^{-1}R(x)$, $(R(x) = \int yp(y|x)dy)$, the cumulative function $F_A(z)$ is $F(x)$, the cumulative function $F_B(z)$ is $\hat{y}^{-1} \int \theta(x' - x')y'dF(x', y')$.

Since statistical inference problems have the same kernel of the integral equations (i.e., the step-function) and the same right-hand side (i.e., the cumulative distribution function), it allows us to introduce (in Section 5) a common standard method (called V-matrix method) for solving all inference problems.

3. Solution of Ill-Posed Operator Equations

In this section, we consider ill-posed operator equations and their solutions.

3.1 Fredholm Integral Equations of the First Kind

In this section, we consider the linear operator equations

$$Af = F, \tag{20}$$

where A maps elements of the metric space $f \in \mathcal{M} \subset E_1$ into elements of the metric space $F \in \mathcal{N} \subset E_2$. Let f be a continuous one-to-one operator and $f(\mathcal{M}) = \mathcal{N}$. Let the solution of such operator equation exist and be unique. Then

$$\mathcal{M} = A^{-1}\mathcal{N}.$$

The crucial question is whether this inverse operator A^{-1} is continuous. If it is, then close functions in $\mathcal{N}$ correspond to close functions in $\mathcal{M}$. That is, "small" changes in the right-hand side of (20) cause "small" changes of its solution. In this case, we call the operator A^{-1} *stable* (Tikhonov and Arsenin, 1977).

If, however, the inverse operator is discontinuous, then "small" changes in the right-hand side of (20) can cause significant changes of the solution. In this case, we call the operator A^{-1} *unstable.*

Solution of equation (20) is called *well-posed* if this solution

1. *exists;*

2. *is unique;*

3. *is stable.*

Otherwise we call the solution *ill-posed.*

We are interested in the situation when the solution of operator equation *exists*, and *is unique.* In this case, the effectiveness of solution of equation (20) is defined by the *stability* of the operator A^{-1}. If the operator is unstable, then, generally speaking, the numerical solution of equation is impossible.

Here we consider linear integral operator

$$Af(x) = \int_a^b K(x,u)f(u)du$$

defined by the kernel $K(t,u)$, which is continuous almost everywhere on $a \leq t \leq b, \quad c \leq x \leq d$. This kernel maps the set of functions $\{f(t)\}$, continuous on $[a,b]$, unto the set of functions $\{F(x)\}$, also continuous on $[c,d]$. The corresponding Fredholm equation of the first kind

$$\int_a^b K(x,u)f(u)du = F(x)$$

requires finding the solution $f(u)$ given the right-hand side $F(x)$.

In this paper, we consider integral equation defined by the so-called convolution kernel

$$K(x,u) = K(x-u).$$

Moreover, we consider the specific convolution kernel of the form

$$K(x-u) = \theta(x-u).$$

As stated in Section 2.2, this kernel covers all settings of empirical inference problems.

First, we show that the solution of equation

$$\int_0^1 \theta(x-u)f(u)du = x \tag{21}$$

is indeed ill-posed[3]. It is easy to check that

$$f(x) = 1$$

is the solution of this equation. Indeed,

$$\int_0^1 \theta(x-u)du = \int_0^x du = x. \tag{22}$$

It is also easy to check that the function

$$f^*(x) = 1 + \cos nx \tag{23}$$

is a solution of the equation

$$\int_0^1 \theta(x-u)f^*(u)du = x + \frac{\sin nx}{n}. \tag{24}$$

That is, when n increases, the right-hand sides of equations (22) and (24) are getting close to each other, but their solutions (21) and (23) are not.

The problem is how one can solve an ill-posed equation when its right-hand side is defined imprecisely.

3.2 Methods of Solving Ill-Posed Problems

In this subsection, we consider methods for solving ill-posed operator equations.

3.2.1 Inverse Operator Lemma

The following classical inverse operator lemma (Tikhonov and Arsenin, 1977) is the key enabler for solving ill-posed problems.

Lemma. *If A is a continuous one-to-one operator defined on a compact set $\mathcal{M}^* \subset \mathcal{M}$, then the inverse operator A^{-1} is continuous on the set $\mathcal{N}^* = A\mathcal{M}^*$.*

Therefore, the conditions of existence and uniqueness of the solution of an operator equation imply that the problem is well-posed on the compact $\mathcal{M}^*$. The third condition (stability of the solution) is automatically satisfied. This lemma is the basis for all constructive ideas of solving ill-posed problems. We now consider one of them.

3.2.2 Regularization Method

Suppose that we have to solve the operator equation

$$Af = F \tag{25}$$

[3] Using the same arguments, one can show that the problem of solving any Fredholm equation of the first kind is ill-posed.

defined by continuous one-to-one operator A mapping $\mathcal{M}$ into $\mathcal{N}$, and assume the solution of (25) exists. Also suppose that, instead of the right-hand side $F(x)$, we are given its approximation $F_\delta(x)$, where

$$\rho_{E_2}(F(x), F_\delta(x)) \leq \delta.$$

Our goal is to find the solution of equation

$$Af = F_\delta$$

when $\delta \to 0$.

Consider a lower semi-continuous functional $W(f)$ (called the *regularizer*) that has the following three properties:

1. the solution of the operator equation (25) belongs to the domain $D(W)$ of the functional $W(f)$;
2. functional $W(f)$ is non-negative values in its domain;
3. all sets
$$\mathcal{M}_c = \{f : W(f) \leq c\}$$
are compact for any $c \geq 0$.

The idea of regularization is to find a solution for (25) as an element minimizing the so-called regularized functional

$$R_\gamma(\hat{f}, F_\delta) = \rho^2_{E_2}(A\hat{f}, F_\delta) + \gamma_\delta W(\hat{f}), \qquad \hat{f} \in D(W) \tag{26}$$

with *regularization parameter* $\gamma_\delta > 0$.

The following theorem holds true (Tikhonov and Arsenin, 1977).

Theorem 1 *Let E_1 and E_2 be metric spaces, and suppose for $F \in \mathcal{N}$ there exists a solution of (25) that belongs to $\mathcal{M}_c$. Suppose that, instead of the exact right-hand side F in (25), its approximations*[4] *$F_\delta \in E_2$ in (26) are given such that $\rho_{E_2}(F, F_\delta) \leq \delta$. Consider the sequence of parameters γ such that*

$$\gamma(\delta) \longrightarrow 0 \;\text{ for }\; \delta \longrightarrow 0,$$
$$\lim_{\delta \longrightarrow 0} \frac{\delta^2}{\gamma(\delta)} \leq r < \infty. \tag{27}$$

Then the sequence of solutions $f_\delta^{\gamma(\delta)}$ minimizing the functionals $R_{\gamma(\delta)}(f, F_\delta)$ on $D(W)$ converges to the exact solution f (in the metric of space E_1) as $\delta \longrightarrow 0$.

In a Hilbert space, the functional $W(f)$ may be chosen as $||f||^2$ for a linear operator A. Although the sets $\mathcal{M}_c$ are (only) weakly compact in this case, regularized solutions converge to the desired one. Such a choice of regularized functional is convenient since its domain $D(W)$ is the whole space E_1. In this case, however, the conditions on the parameters γ are more restrictive than in the case of Theorem 1: namely, γ should converge to zero slower than δ^2.

Thus the following theorem holds true (Tikhonov and Arsenin, 1977).

Theorem 2 *Let E_1 be a Hilbert space and $W(f) = ||f||^2$. Then for $\gamma(\delta)$ satisfying (27) with $r = 0$, the regularized element $f_\delta^{\gamma(\delta)}$ converges to the exact solution f in metric E_1 as $\delta \to 0$.*

[4] The elements F_δ do not have to belong to the set $\mathcal{N}$.

4. Stochastic Ill-Posed Problems

In this section, we consider the problem of solving the operator equation

$$Af = F, \tag{28}$$

where not only its right-hand side is defined approximately ($F_\ell(x)$ instead of $F(x)$), but the operator Af is also defined approximately. Such problem are called *stochastic ill-posed problems.*

In the next subsections, we describe the conditions under which it is possible to solve equation (28), where both the right-hand side and the operator are defined approximately. We first discuss the general theory for solving stochastic ill-posed problems and then consider specific operators describing particular problems, i.e., empirical inference problems described in Sections 2.3, 2.4, and 2.5. For all these problems, the operator has the form

$$A_\ell f = \int \theta(x-u) f(u) dF_\ell(u).$$

We show that rigorous solutions of stochastic ill-posed problem with this operator leverage the so-called V-matrix, which captures some geometric properties of the data; we also describe specific algorithms for solution of our empirical inference problems.

4.1 Regularization of Stochastic Ill-Posed Problems

Consider the problem of solving the operator equation

$$Af = F$$

under the condition where (random) approximations are given not only for the function on the right-hand side of the equation but for the operator as well (*the stochastic ill-posed problem).*

We assume that, instead of the true operator A, we are given a sequence of random continuous operators $A_\ell,\ \ell = 1, 2, ...$ that converges in probability to the operator A (the definition of closeness between two operators will be defined later).

First, we discuss general conditions under which the solution of stochastic ill-posed problem is possible; after that, we consider specific operator equations corresponding to each empirical inference problem.

As before, we consider the problem of solving the operator equation by the regularization method, i.e., by minimizing the functional

$$R^*_{\gamma_\ell}(f, F_\ell, A_\ell) = \rho^2_{E_2}(A_\ell f, F_\ell) + \gamma_\ell W(f). \tag{29}$$

For this functional, there exists a minimum (perhaps, not unique). We define the closeness of operator A and operator A_ℓ as the distance

$$||A_\ell - A|| = \sup_{f \in D} \frac{||A_\ell f - Af||_{E_2}}{W^{1/2}(f)}.$$

The main result for solving stochastic ill-posed problems via regularization method (29) is provided by the following Theorem (Stefanyuk, 1986), (Vapnik, 1998).

Theorem. *For any $\varepsilon > 0$ and any constants $C_1, C_2 > 0$ there exists a value $\gamma_0 > 0$ such that for any $\gamma_\ell \leq \gamma_0$ the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\rho_{E_2}(F_\ell, F) > C_1\sqrt{\gamma_\ell}\} + P\{||A_\ell - A|| > C_2\sqrt{\gamma_\ell}\} \tag{30}$$

holds true.

Corollary. As follows from this theorem, if the approximations $F_\ell(x)$ of the right-hand side of the operator equation converge to the true function $F(x)$ in E_2 with the rate of convergence $r(\ell)$, and the approximations A_ℓ converge to the true operator A in the metric in E_1 defined in (30) with the rate of convergence $r_A(\ell)$, then there exists such a function

$$r_0(\ell) = \max\{r(\ell),\ r_A(\ell)\}; \qquad \lim_{\ell\to\infty} r_0(\ell) = 0,$$

that the sequence of solutions to the equation converges in probability to the true one if

$$\lim_{\ell\to\infty} \frac{r_0(\ell)}{\sqrt{\gamma_\ell}} = 0, \quad \lim_{\ell\to\infty} \gamma_\ell = 0.$$

4.2 Solution of Empirical Inference Problems

In this section, we consider solutions of the integral equation

$$Af = F,$$

where operator A has the form

$$Af = \int \theta(x-u) f(u) dF_1(u),$$

and the right-hand side of the equation is $F_2(x)$. That is, our goal is to solve the integral equation

$$\int \theta(x-u) f(u) dF_1(x) = F_2(x).$$

We consider the case where $F_1(x)$ and $F_2(x)$ are two different cumulative distribution functions. (This integral equation also includes, as a special case, the problem of regression estimation, where $F_2(x) = \int y dP(x,y)$ for non-negative y). This equation defines the main empirical inference problem described in Section 2. The problem of density ratio estimation requires solving this equation when both functions $F_1(x)$ and $F_2(x)$ are unknown but the iid data

$$X_1^1, ..., X_{\ell_1}^1 \sim F_1 \tag{31}$$

$$X_1^1, ..., X_{\ell_2}^1 \sim F_2 \tag{32}$$

are available. In order to solve this equation, we use empirical approximations instead of actual distribution functions, thus obtaining

$$A_{\ell_1} f = \int \theta(x-u) dF_{\ell_1}(u) \tag{33}$$

$$F_{\ell_k}(x) = \frac{1}{\ell_k} \sum_{i=1}^{\ell_k} \theta(x - X_i^k), \quad k = 1, 2,$$

where $F_{\ell_1}(u)$ are and $F_{\ell_2}(x)$ are the empirical distribution functions obtained from data (31) and (32), respectively.

One can show (see (Vapnik, 1998), Section 7.7) that, for sufficiently large ℓ, the inequality

$$||A_\ell - A|| = \sup_f \frac{||A_\ell f - Af||_{E_2}}{W^{1/2}(f)} \leq ||F_\ell - F||_{E_2}$$

holds true for the smooth solution $f(x)$ of our equations.

From this inequality, bounds (4), and the Theorem of Section 4.1, it follows that the regularized solutions of our operator equations converge to the actual function

$$\rho_{E_1}(f_\ell, f) \to_{\ell \to \infty} 0$$

with probability one.

Therefore, to solve our inference problems, we minimize the functional

$$R_\gamma(f, F_\ell, A_{\ell_1}) = \rho^2_{E_2}(A_{\ell_1} f, F_{\ell_2}) + \gamma_\ell W(f). \tag{34}$$

In order to do this well and find the unique solution of this problem, we have to define three elements of (34):

1. The distance $\rho_{E_2}(F_1, F_2)$ between functions $F_1(x)$ and $F_2(x)$ in E_2.

2. The regularization functional $W(f)$ in the space of functions $f \in E_1$.

3. The rule for selecting the regularization constant γ_ℓ.

In the next sections, we consider the first two elements.

5. Solving Statistical Inference Problems with V-matrix

Consider the explicit form of the functional for solving our inference problems. In order to do this, we specify expressions for the squared distance and regularization functional in expression (34).

5.1 The V-matrix

In this subsection, we consider the key element of our approach, the V-matrix.

5.1.1 Definition of Distance

Let our distance in E_2 be defined by the L_2 metric

$$\rho^2_{E_2}(F_1(x), F_2(x)) = \int (F_1(x) - F_2(x))^2 \sigma(x) d\mu(x),$$

where $\sigma(x)$ is a *known* positive function and $\mu(x)$ is a *known* measure defined on $\mathcal{X}$. To define distance, one can use any non-negative measurable function $\sigma(x)$ and any measure

$\mu(x)$. For example, if our equation is defined in the box domain $[0,1]^d$, we can use uniform measure in this domain and $\sigma(x)=1$.

Below we define the measure $\mu(x)$ as

$$d\mu(x)=\prod_{k=1}^{d} dF_\ell(x^k), \tag{35}$$

where each $F_\ell(x^k)$ is the marginal empirical cumulative distribution function of the coordinate x^k estimated from data.

We also choose function $\sigma(x)$ in the form

$$\sigma(x)=\prod_{k=1}^{n} \sigma_k(x^k). \tag{36}$$

In this paper, we consider several weight functions $\sigma(x^k)$:

1. The function
$$\sigma(x^k)=1.$$

2. For the problem of conditional probability estimation, we consider the function
$$\sigma(x^k)=\frac{1}{F_\ell(x^k|y=1)(1-F_\ell(x^k|y=1))+\epsilon}, \tag{37}$$
where $\varepsilon>0$ is a small constant.

3. For the problem of regression estimation, we consider the case where $y\geq 0$ and, instead of $F_\ell(x^k|y=1)$ in (37), the monotonic function
$$F_\ell(x^k)=\frac{1}{\ell\hat{y}_{av}}\sum_{i=1}^{\ell} y_i\theta(x^k-X_i^k)$$
is used, where $\hat{y}_{av}$ is the average value of y in the training data. This function has properties of ECDF.

4. For the problem of density ratio estimation, we consider an estimate of function $F_{\text{num}}(x)$ instead of the estimate of function $F(x|y=1)$ in (37).

Remark. In order to explain choice (37) for function $\sigma(x)$, consider the problem of one-dimensional conditional probability estimation. Let $f_0(x)$ be the true conditional probability. Consider the function $\hat{f}_0(x)=p_1f_0(x)$. Then the solution of integral equation

$$\int \theta(x-u)\hat{f}(u)dF(u)=F(x|y=1)$$

defines the conditional probability $\hat{f}_0(x)=p_1f_0(x)$. Consider two functions: the estimate of the right-hand side of equation $F_\ell(x|y=1)$ and the actual right-hand side

$$F_0(x|y=1)=\int_{-\infty}^{x}\hat{f}_0(t)dt.$$

The deviation

$$\Delta = F_0(x|y=1) - F_\ell(x|y=1)$$

between these two functions has different values of variance for different x. The variance is small (equal to zero) at the end points of an interval and is large somewhere inside it. To obtain the *uniform relative deviation* of approximation from the actual function over the whole interval, we adjust the distance in any point of interval proportionally to the inverse of variance. Since for any fixed x the variance is

$$\mathrm{Var}(x) = F(x|y=1)(1 - F(x|y=1)), \tag{38}$$

we normalize the squared deviation Δ^2 by (38). The expression (37) realizes this idea.

5.1.2 Definition of Distance for Conditional Probability Estimation Problem

Consider the problem of conditional probability estimation.

For this problem, the squared distance between approximations of the right-hand side and the left-hand side of equation

$$F_\ell(x, y=1) = p_\ell F_\ell(x|y=1) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \theta(x - X_i)$$

can be written as follows:

$$\rho^2(A_\ell f, F_\ell) = \int \left(\int \theta(x-u) f(u) dF_\ell(u) - \int y_i \theta(x-u) dF_\ell(u) \right)^2 \sigma(x) d\mu(x),$$

where $y_i \in \{0, 1\}$ and $F_\ell(x)$ is the empirical distribution function estimated from training vectors X_i. Therefore, we obtain the expression

$$\begin{aligned} \rho^2(A_\ell f, F_\ell) = \quad & \frac{1}{\ell^2}\sum_{i,j=1}^{\ell} f(X_i) f(X_j) \int \theta(x - X_i)\theta(x - X_j)\sigma(x) d\mu(x) - \\ & \frac{2}{\ell^2}\sum_{i,j=1}^{\ell} f(X_i) y_j \int \theta(x - X_i)\theta(x - X_j)\sigma(x) d\mu(x) + \\ & \frac{1}{\ell^2}\sum_{i,j=1}^{\ell} y_i y_j \int \theta(x - X_i)\theta(x - X_j)\sigma(x) d\mu(x), \end{aligned} \tag{39}$$

where the last term does not depend on function $f(x)$.

Since both $\sigma(x)$ and $\mu(x)$ are products of one-dimensional functions, each integral in (39) has the form

$$V_{i,j} = \int \theta(x - X_i)\theta(x - X_j)\sigma(x)\, d\mu(x) = \prod_{k=1}^{d} \int \theta(x^k - X_i^k)\theta(x^k - X_j^k)\sigma_k(x^k) d\mu(x^k). \tag{40}$$

This $(\ell \times \ell)$-dimensional matrix of elements $V_{i,j}$ we call V-matrix of the sample $X_1, ..., X_\ell$, where $X_i = (X_i^1, \ldots X_i^d), \quad \forall i = 1, \ldots, \ell$.

Consider three cases:

Case 1. Data belongs to the upper-bounded support $(-\infty, B]^d$ for some B and $\sigma(x) = 1$ on this support. Then the elements $V_{i,j}$ of V-matrix have the form

$$V_{i,j} = \prod_{k=1}^{d} (B - \max\{X_i^k, X_j^k\}).$$

Case 2. Case where $\sigma(x^k) = 1$ and μ defined as (35). Then the elements $V_{i,j}$ of V-matrix have the form

$$V_{i,j} = \prod_{k=1}^{d} \nu(X^k > \max\{X_i^k, X_j^k\}).$$

where $\nu(X^k > \max\{X_i^k, X_j^k\})$ is the frequency of X^k from the given data with the values larger than $\max\{X_i^k, X_j^k\}$.

Case 3. Case where $\sigma(x)$ is defined as (36), (37) and $\mu(x)$ as (35). In this case, the values $V_{i,j}$ also can be easily computed numerically (since both functions are piecewise constant, the integration (40) is reduced to a summation of constants).

To rewrite the expression for the distance in a compact form, we introduce the ℓ-dimensional vector Φ

$$\Phi = (f(X_1), ..., f(X_\ell))^T.$$

Then, taking into account (39), we rewrite the first summand of functional (34) as

$$\rho^2(A_\ell f, F_\ell) = \frac{1}{\ell^2}\left(\Phi^T V \Phi - 2\Phi^T V Y + Y^T V Y\right), \tag{41}$$

where Y denotes the ℓ-dimensional vector $(y_1, ..., y_\ell)^T$, $y_i \in \{0, 1\}$.

5.1.3 Distance for Regression Estimation Problem

Repeating the same derivation for regression estimation problem, we obtain the same expression for the distance

$$\rho^2(A_\ell f, F_\ell) = \frac{1}{\ell^2}\left(\Phi^T V \Phi - 2\Phi^T V Y + Y^T V Y\right),$$

where coordinates of vector Y are values $y \in R^1$ given in examples (17) for regression estimation problem.

5.1.4 Distance for Density Ratio Estimation Problem

In the problem of density ratio estimation, we have to solve the integral equation

$$\int \theta(x - u) R(u) dF_{\text{den}}(u) = F_{\text{num}}(x),$$

where cumulative distribution functions $F_{\text{den}}(x)$ and $F_{\text{num}}(x)$ are unknown but iid data

$$X_1, ..., X_{\ell_{\text{den}}} \sim F_{\text{den}}(x)$$

and iid data

$$X'_1, ..., X'_{\ell_{\text{num}}} \sim F_{\text{num}}(x)$$

are available.

Using the empirical estimates

$$F_{\ell_{\text{num}}}(x) = \frac{1}{\ell_{\text{num}}} \sum_{j=1}^{\ell_{\text{num}}} \theta(x - X'_j)$$

and

$$F_{\ell_{\text{den}}}(x) = \frac{1}{\ell_{\text{den}}} \sum_{i=1}^{\ell_{\text{den}}} \theta(x - X_i)$$

instead of unknown cumulative distribution $F_{\text{num}}(x)$ and $F_{\text{den}}(x)$ and repeating the same distance computations as in the problems of conditional probability estimation and regression estimation, we obtain

$$\begin{aligned}\rho^2 = & \int \left(\int \theta(x-u) R(u) dF_{\ell_{\text{den}}}(u) - F_{\ell_{\text{num}}}(x) \right)^2 \sigma(x) d\mu(x) = \\ & \frac{1}{\ell_{\text{den}}^2} \sum_{i,j=1}^{\ell_{\text{den}}} R(X_i) R(X_j) \int \theta(x - X_j) \theta(x - X_j) \sigma(x) d\mu(x) - \\ & \frac{2}{\ell_{\text{num}} \ell_{\text{den}}} \sum_{i=1}^{\ell_{\text{num}}} \sum_{j=1}^{\ell_{\text{den}}} R(X_i) R(X'_j) \int \theta(x - X_i) \theta(x - X'_j) \sigma(x) d\mu(x) + \\ & \frac{1}{\ell_{\text{num}}^2} \sum_{i,j=1}^{\ell_{\text{num}}} \int \theta(x - X'_j) \theta(x - X'_j) \sigma(x) d\mu(x) = \frac{1}{\ell_{\text{num}}^2} \sum_{i,j=1}^{\ell_{\text{num}}} V^{**}_{i,j} + \\ & \frac{1}{\ell_{\text{den}}^2} \sum_{i,j=1}^{\ell_{\text{den}}} R(X_i) R(X_j) V_{i,j} - \frac{2}{\ell_{\text{num}} \ell_{\text{den}}} \sum_{i=1}^{\ell_{\text{den}}} \sum_{j=1}^{\ell_{\text{num}}} R(X_i) R(X'_j) V^*_{i,j},\end{aligned}$$

where the values $V_{i,j}, V^*_{i,j}, V^{**}_{i,j}$ are calculated as

$$\begin{cases} V_{i,j} = & \int \theta(x - X_i) \theta(x - X_j) \sigma(x) d\mu(x), \quad i, j = 1, ..., \ell_{\text{den}}, \\ V^*_{i,j} = & \int \theta(x - X_i) \theta(x - X'_j) \sigma(x) d\mu(x), \quad i = 1, ..., \ell_{\text{num}}, \ j = 1, ..., \ell_{\text{den}}, \\ V^{**}_{i,j} = & \int \theta(x - X'_i) \theta(x - X'_j) \sigma(x) d\mu(x), \quad i, j = 1, ..., \ell_{\text{num}}. \end{cases}$$

We denote by V, V^*, and V^{**} (respectively, $(\ell_{\text{den}} \times \ell_{\text{den}})$-dimensional, $(\ell_{\text{den}} \times \ell_{\text{num}})$-dimensional, and $(\ell_{\text{num}} \times \ell_{\text{num}})$-dimensional) the matrices of corresponding elements $V_{i,j}$, $V^*_{i,j}$, and $V^{**}_{i,j}$. We also denote by $1_{\ell_{\text{num}}}$ the ℓ_{num}-dimensional vector of ones, and by R – the ℓ_{den}-dimensional column vector of $R(X_i)$, $i = 1, \ldots, \ell_{\text{den}}$.

Using these notations, we can rewrite the distance as follows:

$$\rho^2 = \frac{1}{\ell_{\text{den}}^2} \left(R^T V R - 2 \left(\frac{\ell_{\text{den}}}{\ell_{\text{num}}} \right) R^T V^* 1_{\ell_{\text{num}}} + \left(\frac{\ell_{\text{den}}}{\ell_{\text{num}}} \right)^2 1^T_{\ell_{\text{num}}} V^{**} 1_{\ell_{\text{num}}} \right).$$

5.2 The Regularization Functionals in RKHS

For each of our inference problems, we now look for its solution in Reproducing Kernel Hilbert Space (RKHS).

5.2.1 Reproducing Kernel Hilbert Space

According to Mercer theorem, any positive semi-definite kernel has a representation

$$K(x,z)=\sum_{k=1}^{\infty}\lambda_k\phi_k(x)\phi_k(z), \quad x,z\in\mathcal{X},$$

where $\{\phi_k(x)\}$ is a system of orthonormal functions and $\lambda_k\ge 0 \ \ \forall k$.

Consider the set of functions

$$f(x;a)=\sum_{k=1}^{\infty}a_k\phi_k(x). \tag{42}$$

We say that set of functions (42) belongs to RKHS of kernel $K(x,z)$ if we can define the inner product (f_1,f_2) in this space such that

$$(f_1(x),K(x,y))=f_1(y). \tag{43}$$

It is easy to check that the inner product

$$(f(x,a),f(x,b))=\sum_{k=1}^{\infty}\frac{a_kb_k}{\lambda_k},$$

where a_k and b_k are the coefficients of expansion of functions $f(x,a)$, and $f(x,b)$, satisfies the reproducing property (43). In particular, the equality

$$(K(x_1,z),K(x_2,z))=K(x_1,x_2) \tag{44}$$

holds true for the kernel $K(x,x^*)$ that defines RKHS.

The remarkable property of RKHS is the so-called Representer Theorem (Kimeldorf and Wahba, 1971), (Kimeldorf and Wahba, 1970), (Schölkopf et al., 2001), which states that any function $f(x)$ from RKHS that minimizes functional (34) can be represented as

$$f(x)=\sum_{i=1}^{\ell}c_iK(X_i,x),$$

where $c_i, \ i=1,...,\ell$ are parameters and $X_i, \ i=1,...,\ell$ are vectors of observations.

5.2.2 Explicit Form of Regularization Functional.

In all our Statistical Inference problems, we are looking for solutions in RKHS, where we use the squared norm as the regularization functional:

$$W(f)=(f,f)=||f||^2. \tag{45}$$

That is, we are looking for solution in the form

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(X_i, x), \tag{46}$$

where X_i are elements of the observation. Using property (44), we define the functional (45) as

$$W(f) = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j).$$

In order to use the matrix form of (34), we introduce the following notations:

1. K is the $(\ell \times \ell)$-dimensional matrix of elements $K(X_i, X_j),\ i, j = 1, ..., \ell$.
2. $\mathcal{K}(x)$ is the ℓ-dimensional vector of functions $K(X_i, x),\quad i = 1, ..., \ell$.
3. A is the ℓ-dimensional vector $A = (\alpha_1, ..., \alpha_\ell)^T$ of elements $\alpha_i,\quad i = 1, ..., \ell$.

In these notations, the regularization functional has the form

$$W(f) = A^T K A, \tag{47}$$

and its solution has the form

$$f(x) = A^T \mathcal{K}(x). \tag{48}$$

6. Solution of Statistical Inference Problems

In this section, we formulate our statistical inference problems as optimization problems.

6.1 Estimation of Conditional Probability Function

Here we present an explicit form of the optimization problem for estimating conditional probability function.

We are looking for the solution in form (48), where we have to find vector A. In order to find it, we have to minimize the objective function

$$T(A) = A^T K V K A - 2A^T K V Y + \gamma A^T K A, \tag{49}$$

where Y is a binary vector (with coordinates $y \in \{0, 1\}$) defined by the observations. The first two terms of the objective function come from distance (41), the last term is regularization functional (47). (The third term from (49) was omitted in the target functional since it does not depend on the unknown function.) Since the conditional probability has values between 0 and 1, we have to minimize this objective function subject to the constraint

$$0 \leq A^T \mathcal{K}(x) \leq 1, \quad \forall x \in X. \tag{50}$$

We also know that

$$\int A^T \mathcal{K}(x) dF(x) = p_0, \tag{51}$$

where p_0 is the probability of class $y = 1$.

Minimization of (49) subject to constraints (50), (51) is a difficult optimization problem. To simplify this problem, we minimize the functional subject to the constraints

$$0 \leq A^T \mathcal{K}(X_i) \leq 1, \ \ i = 1, ..., \ell, \tag{52}$$

defined only at the vectors X_i of observations[5].

Also, we can approximate equality (51) using training data

$$\frac{1}{\ell} \sum_{i=1}^{\ell} A^T \mathcal{K}(X_i) = p_\ell, \tag{53}$$

where p_ℓ is the frequency of class $y = 1$ estimated from data. Using matrix notation, the constraints (52) and (53) can be rewritten as follows:

$$0_\ell \leq KA \leq 1_\ell, \tag{54}$$

$$\frac{1}{\ell} A^T K 1_\ell = p_\ell. \tag{55}$$

where K is the matrix of elements $K(X_i, X_j), \ i, j = 1, ..., \ell$ and 0_ℓ, 1_ℓ are ℓ-dimensional vectors of zeros and ones, respectively.

Therefore, we are looking for the solution in form (48), where parameters of vector A minimize functional (49) subject to constraints (54) and (55). This is a quadratic optimization problem with one linear equality constraint and 2ℓ *general* linear inequality constraints.

In Section 6.4, we simplify this optimization problem by reducing it to a quadratic optimization problem with one linear equality constraint and several *box* constraints.

6.2 Estimation of Regression Function

Similarly, we can formulate the problem of regression function estimation, which has the form (48). To find the vector A, we minimize the functional

$$T(A) = A^T KVKA - 2A^T KVY + \gamma A^T KA, \tag{56}$$

where Y is a real-valued vector (with coordinates $y_i \in R^1$ of observations (5)).

Suppose that we have the following knowledge about the regression function:

1. Regression $y = f(x) = A^T \mathcal{K}(x)$ takes values inside an interval $[a, b]$:

$$a \leq A^T \mathcal{K}(x) \leq b, \quad \forall x \in \mathcal{X}. \tag{57}$$

2. We know the expectation of the values of the regression function:

$$\int A^T \mathcal{K}(x) dF(x) = c. \tag{58}$$

[5] One can find the solution in closed form $A = (VK + \gamma I)^{-1} VY$ if constraints (52), (53) are ignored; here I is the identity matrix.

Then we can solve the following problem: minimize functional (56) subject to constraints (57), (58).

Usually we do not have knowledge (57), (58), but we can approximate it from the training data. Specifically, we can approximate a by the smallest value a_ℓ of y_i, while b can be approximated by the largest value b_ℓ of y_i from the training set:

$$a_\ell = \min\{y_1, ..., y_\ell\}, \quad b_\ell = \max\{y_1, ..., y_\ell\}.$$

We then consider constraint (57) applied only for the training data:

$$a_\ell \leq A^T \mathcal{K}(X_i) \leq b_\ell, \quad i = 1, ..., \ell. \tag{59}$$

Also, we can approximate (58) with the equality constraint

$$\frac{1}{\ell} \sum_{i=1}^{\ell} A^T \mathcal{K}(X_i) = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i. \tag{60}$$

Constraints (59), (60) can be written in matrix notation

$$a_\ell 1_\ell \leq KA \leq 1_\ell b_\ell, \tag{61}$$

$$\frac{1}{\ell} A^T K 1_\ell = \hat{y}_{av}, \tag{62}$$

where $\hat{y}_{av}$ is the right-hand side of (60). If these approximations[6] are reasonable, the problem of estimating the regression can be stated as minimization of functional (56) subject to constraints (61), (62). This is a quadratic optimization problem with one linear equality constraint and 2ℓ general linear inequality constraints.

6.3 Estimation of Density Ratio Function

To solve the problem of estimating density ratio function in the form

$$R(x) = A^T \mathcal{K}(x),$$

where A is the ℓ_{den}-dimensional vector of parameters and $\mathcal{K}(x)$ is the ℓ_{den}-dimensional vector of functions $K(X_1, x), ..., K(X_{\ell_{\text{den}}}, x)$, we have to minimize the functional

$$T(A) = A^T KVKA - 2 \left(\frac{\ell_{\text{den}}}{\ell_{\text{num}}} \right) A^T KV^* \mathbf{1}_{\ell_{\text{num}}} + \gamma A^T KA, \tag{63}$$

where K is the $(\ell_{\text{den}} \times \ell_{\text{den}})$-dimensional matrix of elements $K(X_i, X_j)$ subject to the constraints

$$A^T \mathcal{K}(x) \geq 0, \quad \forall x \in X,$$

$$\int A^T \mathcal{K}(x) dF_{\text{den}}(x) = 1.$$

[6] Without constraints, the solution has the closed form (see footnote 5), where $y \in R^1$ are elements of training data for regression.

As above, we replace these constraints with their approximations

$$KA \geq \mathbf{0}_{\ell_{\text{den}}},$$

$$\frac{1}{\ell_{\text{den}}} A^T K V^* 1_{\ell_{\text{num}}} = 1.$$

Here K is $(\ell_{\text{den}} \times \ell_{\text{den}})$-dimensional matrix of observations from $F_{\text{den}}(x)$, and V^* is $(\ell_{\text{den}} \times \ell_{\text{num}})$-dimensional matrix defined in Section 5.1.

6.4 Two-Stage Method for Function Estimation: Data Smoothing and Data Interpolation

Solutions of Statistical Inference problems considered in the previous sections require numerical treatment of the general quadratic optimization problem: minimization of quadratic form subject to one linear equality constraint and 2ℓ linear inequality constraints of general type (ℓ linear inequality constraints for density ratio estimation problem).

Numerical solution for such problems can be computationally hard (especially when ℓ is large). In this section, we simplify the problem by splitting it into two stages:

1. Estimating function values at ℓ observation points, that is, the estimating vector $\Phi = (f(X_1), ..., f(X_\ell))^T$.
2. Interpolating the values of function known at the ℓ observation points to other points in the space $\mathcal{X}$.

6.4.1 Stage 1: Estimating Function Values at Observation Points

In order to find the function values at the training data points, we rewrite the regularization functional in objective functions (49), (56), (63) in a different form. In order to do this, we use the equality

$$K = KK^+K,$$

where K^+ is the *generalized inverse* matrix of matrix[7] K.

In our regularization term of objective functions, we use the equality

$$A^T K A = A^T K K^+ K A.$$

1. Estimation of values of conditional probability. For the problem of estimating the values of conditional probability at ℓ observation points, we rewrite objective function (49) in the form

$$W(\Phi) = \Phi^T V \Phi - 2\Phi^T V Y + \gamma \Phi^T K^+ \Phi, \tag{64}$$

where we have denoted

$$\Phi = KA. \tag{65}$$

In the problem of estimating conditional probability, Y is a binary vector.

[7] Along with generalized inverse matrix, pseudoinverse matrix is also used. Pseudoinverse matrix M^+ of the matrix M (not necessarily symmetric) satisfies the following four conditions: (1) $MM^+M = M$, (2) $M^+MM^+ = M^+$, (3) $(MM^+)^T = MM^+$, and (4) $(M^+M)^T = M^+M$. If matrix M is invertible, then $M^+ = M^{-1}$. Pseudoinverse exists and is unique for any matrix.

In order to find vector Φ, we minimize functional (64) subject to box constraints

$$\mathbf{0}_\ell \leq \Phi \leq 1_\ell,$$

and equality constraint

$$\frac{1}{\ell}\Phi^T 1_\ell = p_\ell.$$

2. Estimating values of regression. In order to estimate the vector Φ of values of regression at ℓ observation points, we minimize functional (64) (where Y is a real-valued vector), subject to the box constraints

$$a_\ell 1_\ell \leq \Phi \leq b_\ell 1_\ell,$$

and the equality constraint

$$\frac{1}{\ell}\Phi^T 1_\ell = \hat{y}_{av}.$$

3. Estimating values of density ratio function. In order to estimate the vector Φ of values of density ratio function at ℓ_{den} observation points $X_1, ..., X_{\ell_{\text{den}}}$, we minimize the functional

$$\Phi^T V \Phi - 2\left(\frac{\ell_{\text{den}}}{\ell_{\text{num}}}\right)\Phi^T V^* \mathbf{1}_{\ell_{\text{num}}} + \gamma \Phi^T K^+ \Phi$$

subject to the box constraints

$$\Phi \geq \mathbf{0}_{\ell_{\text{den}}},$$

and the equality constraint

$$\frac{1}{\ell_{\text{den}}}\Phi^T V^* 1_{\ell_{\text{num}}} = 1.$$

6.4.2 Stage 2: Function Interpolation

In the second stage of our two-stage procedure, we use the estimated function values at the points of training set to define the function in input space. That is, we solve the problem of function interpolation.

In order to do this, consider representation (65) of vector Φ^*:

$$\Phi^* = KA^*. \tag{66}$$

We also consider the RKHS representation of the desired function:

$$f(x) = A^{*T}\mathcal{K}(x). \tag{67}$$

If the inverse matrix K^{-1} exists, then

$$A^* = K^{-1}\Phi^*.$$

If K^{-1} does not exist, there are many different A^* satisfying (66). In this situation, the best interpolation of Φ^* is a (linear) function (67) that belongs to the subset of functions with the smallest bound on VC dimension (Vapnik, 1998). According to Theorem 10.6

in (Vapnik, 1998), such a function either satisfies equation (66) with the smallest L_2 norm of A^* or it satisfies equation (66) with the smallest L_0 norm of A^*.

Efficient computational implementations for both L_0 and L_2 norms are available in the popular scientific software package Matlab.

Note that the obtained solutions in all our problems satisfy the corresponding constraints only on the training data, but they do not have to satisfy these constraints at any $x \in \mathcal{X}$. Therefore, we truncate the obtained solution functions as

$$f_{tr}(x) = [A^{*T}\mathcal{K}(x)]_a^b,$$

where

$$[u]_a^b = \begin{cases} a, & \text{if } u < a \\ u, & \text{if } a \leq u \leq b \\ b, & \text{if } u > b \end{cases}$$

Remark. For conditional probability estimation, the choice of $a > 0$, $b < 1$ (for constraints in training and truncation in test) is an additional tool for regularization that can leverage prior knowledge.

6.4.3 Additional Considerations

For many problems, it is useful to consider the solutions in the form of a function from a set of RKHS functions with a bias term:

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(X_i, x) + c = A^T \mathcal{K}(x) + c.$$

Using this set of functions, our quadratic optimization formulation for estimating the function values at training data points for the problem of conditional probability and regression estimation is as follows: minimize the functional (over vectors Φ)

$$(\Phi + c1_\ell)^T V (\Phi + c1_\ell) - 2(\Phi + c1_\ell)^T VY + \gamma \Phi^T K^+ \Phi$$

subject to the constraints

$$(a - c1_\ell) \leq \Phi \leq (b - c1_\ell),$$

(where $a = 0,\ b = 1$ for conditional probability problem, and $a = a_\ell,\ b = b_\ell$ for regression problem).

$$\frac{1}{\ell} 1_\ell^T + c = \hat{y}_{av}$$

where we denoted

$$\hat{y}_{av} = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i.$$

For estimating the values of density ratio function at points $(X_1, \ldots, X_{\ell_{\text{den}}})$, we minimize the functional

$$(\Phi + c1_{\ell_{\text{den}}})^T V (\Phi + c1_{\ell_{\text{den}}}) - 2\left(\frac{\ell_{\text{den}}}{\ell_{\text{num}}}\right)(\Phi + c1_{\ell_{\text{den}}})^T V^* 1_{\ell_{\text{num}}} + \gamma \Phi^T K^+ \Phi$$

subject to the constraints

$$-c1_{\ell_{\text{den}}} \leq \Phi,$$
$$\Phi^T 1_{\ell_{\text{den}}} + c\ell_{\text{den}} = \ell_{\text{den}}.$$

7. Applications of Density Ratio Estimation

Here we describe three applications of density ratio estimation (Sugiyama et al., 2012), (Kawahara and Sugiyama, 2009), specifically,

– Data adaptation and correction of solution for unbalanced data.

– Estimation of mutual information and problem of feature selection.

– Change point detection.

It is important to note that, in all these problems, it is required to estimate not the function $R(x)$, but rather the values $R(X_i)$ of density ratio function at the points $X_1, ..., X_{\ell_{\text{den}}}$ (generated by probability measure $F_{\text{den}}(x)$).

Below we consider the first two problems in the pattern recognition setting and then consider two new applications:
1) Learning from data with unbalanced classes
2) Learning of local rules.

7.1 Data Adaptation Problem

Let the iid data

$$(y_1, X_1), ..., (y_\ell, X_\ell) \tag{68}$$

be defined by a fixed unknown density function $p(x)$ and a fixed unknown conditional density function $p(y|x)$ generated according to an unknown joint density function $p(x, y) = p(y|x)p(x)$. Suppose now that one is given data

$$X_1^*, ..., X_{\ell_1}^* \tag{69}$$

defined by another fixed unknown density function $p_*(x)$. This density function, together with conditional density $p(y|x)$ (the same one as for Equation 68), defines the joint density function $p_*(x, y) = p(y|x)p_*(x)$.

It is required, using data (68) and (69), to find in a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one that minimizes the functional

$$T(\alpha) = \int L(y, f(x, \alpha))p_*(x, y)dydx, \tag{70}$$

where $L(\cdot, \cdot)$ is a known loss function.

This setting is an important generalization of the classical function estimation problem where the functional dependency between variables y and x is the same (the function $p(y|x)$ which is the part of composition of $p(x, y)$ and $p_*(x, y)$), but the environments (defined by densities $p(x)$ and $p_*(x)$) are different.

It is required, by observing examples from one environment (with $p(x)$), to define the rule for another environment (with $p^*(x)$). Let us denote

$$R(x) = \frac{p_*(x)}{p(x)}, \quad p(x) > 0.$$

Then functional (70) can be rewritten as

$$T(\alpha) = \int L(y, f(x, \alpha))R(x)p(x, y)dydx,$$

and we have to minimize the functional

$$T_\ell(\alpha) = \sum_{i=1}^{\ell} L(y_i, f(X_i, \alpha))R(x_i),$$

where $X_i,\ y_i$ are data points from (68). In this equation, we have multipliers $R(X_i)$ that define the adaptation of data (69) generated by joint density $p(x, y) = p(y|x)p(x)$ to the data generated by the density $p_*(x, y) = p(y|x)p_*(x)$. Knowledge of density ratio values $R(X_i)$ leads to a modification of classical algorithms.

For SVM method in pattern recognition (Vapnik, 1995), (Vapnik, 1998), this means that we have to minimize the functional

$$T_\ell(w) = (w, w) + C \sum_{i=1}^{\ell} R(X_i)\xi_i \tag{71}$$

(C is a tuning parameter) subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i, \quad \xi \geq 0, \quad y_i \in \{-1, +1\}, \tag{72}$$

where z_i is the image of vector $X_i \in \mathcal{X}$ in feature space $\mathcal{Z}$.

This leads to the following dual-space SVM solution: maximize the functional

$$T_\ell(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(X_i, X_j), \tag{73}$$

where $(z_i, z_j) = K(X_i, X_j)$ is Mercer kernel that defines the inner product (z_i, z_j) subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{74}$$

and the constraints

$$0 \leq \alpha_i \leq CR(X_i). \tag{75}$$

The adaptation to new data is given by the values $R(x_i),\ i = 1, ..., \ell$; these values are set to 1 in standard SVM (71).

7.2 Estimation of Mutual Information.

Consider k-class pattern recognition problem $y \in \{a_1, ..., a_k\}$.

The *entropy* of nominal random variable y (level of uncertainty for y with no information about corresponding x) is defined by

$$H(y) = -\sum_{t=1}^{k} p(y = a_t) \log_2 p(y = a_t).$$

Similarly, the *conditional entropy* given fixed value x_* (level of uncertainty of y given information x_*) is defined by the value

$$H(y|x_*) = -\sum_{t=1}^{k} p(y = a_t|x_*) \log_2 p(y = a_t|x_*).$$

For any x^*, the difference (decrease in uncertainty)

$$\Delta H(y|x_*) = H(y) - H(y|x_*)$$

defines the amount of information about y contained in vector x_*. The expectation of this value (with respect to x)

$$I(x, y) = \int \Delta H(y|x) dF(x)$$

is called the *mutual information* between variables y and vectors x. It describes how much information vector x caries about variable y. The mutual information can be rewritten in the form

$$I(x, y) = \sum_{t=1}^{k} p(y = a_t) \int \left(p(x, y = a_t) \log_2 \frac{p(x, y = a_t)}{p(x)p(y = a_t)} \right) dF(x) \tag{76}$$

(see (Cover and Thomas, 2006) for details).

For two densities ($p(x|y = a_t)$ and $p(x)$), the density ratio function is

$$R(x, y = a_t) = \frac{p(x|y = a_t)}{p(x)}.$$

Using this notation, one can rewrite expression (76) as

$$I(x, y) = \sum_{t=1}^{k} p(y = a_t) \int R(y = a_t, x) \log_2 R(y = a_t, x) dF(x), \tag{77}$$

where $F(x)$ is cumulative distribution function of x.

Our goal is to use data

$$(y_1, X_1), ..., (y_\ell, X_\ell)$$

to estimate $I(x, y)$. Using in (77) the empirical distribution function $F_\ell(x)$ and the values $p_\ell(y = a_t)$ estimated from the data, we obtain the approximation $I_\ell(x, y)$ of mutual information (77):

$$I_\ell(x, y) = \frac{1}{\ell} \sum_{t=1}^{m} p(y = a_t) \sum_{i=1}^{\ell} R(X_i, y = a_t) \log_2 R(X_i, y = a_t).$$

Therefore, in order to estimate the mutual information for k-class classification problem, one has to solve the problem of values of density ratio estimation problem k times at the observation points $R(X_i, y = a_t),\ i = 1, ..., \ell$ and use these values in (77).

Feature selection problem and mutual information. Estimates of mutual information play important role in the problem of feature selection. Indeed, the problem of selecting k features from the set of n features require to find among n features $x^1,..,x^n$ such k elements $x^{k_1},...,x^{k_k}$ which contain *maximal information* about variable y generated according to $p(y|x),\ x=(x^1,...,x^n)$. That means to find the subset of k elements with maximal mutual information. This is a hard computational problem: even if one can estimate the mutual information from data well, one still needs to solve mutual information estimation problem C_n^k times to chose the best subset.

Therefore some heuristic methods are used (Brown et al., 2012) to chose the subset with best features. There are two heuristic approaches to the problem of estimating best features:

1. To estimate mutual information $I(y,x^t)$ or $I(x^t,x^m)$ of scalar values and then combine (heuristically) the results.

2. To estimate mutual information between the value of y and two features $x_{t,m}=(x^t,x^m)$, obtaining n^2 elements of matrix $I(y,x_{t,m})\ t,m=1,\dots,n$ and choose from this matrix the minor with the largest score (say, the sum of its elements).

All these procedures require accurate estimates of mutual information.

7.3 Unbalanced Classes in Pattern Recognition

An important application of data adaptation method is the case of binary classification problem with unbalanced training data (du Plessis and Sugiyama, 2012). In this case, the numbers of training examples for both classes differ significantly (often, by orders of magnitude). For instance, for diagnosis of rare diseases, the number of samples from the first class (patients suffering from the disease) is much smaller than the number of samples from the second class (patients without that disease).

Classical pattern recognition algorithms applied to unbalanced data can lead to large false positive or false negative error rates. We would like to construct a method that would allow to control the balance of both error rates. Formally, this means that training data are generated according to some probability measure

$$p(x)=p(x|y=1)p+p(x|y=0)(1-p),$$

where $0\le p\le 1$ is a fixed parameter that defines probability of the event of the first class. Learning algorithms are developed to minimize the expectation of error for this generator of random events.

Our goal, however, is to minimize the expected error for another generator

$$p_*(x)=p(x|y=1)p_*+p(x|y=0)(1-p_*),$$

where p_* defines different probability of the first class (in the rare disease example, we minimize the expected error if this disease is not so rare); that is, for parameter $p=p_*$.

To solve this problem, we have to estimate the values of density ratio function

$$R(x)=\frac{p_*(x)}{p(x)}$$

from available data. Suppose we are given observations

$$(y_1, X_1), ..., (y_\ell, X_\ell). \tag{78}$$

Let us denote by X_i^1 and X_j^0 vectors from (78) corresponding to $y = 1$ and $y = 0$, respectively. We rewrite elements of x from (78) generated by $p(x)$ as

$$X_{i_1}^1, ..., X_{i_m}^1, X_{i_{m+1}}^0, ..., X_{i_\ell}^0$$

Consider the new training set that imitates iid observations generated by $p_*(x)$ by having the elements of the first class to have frequency p_*:

$$X_{i_1}^1, ..., X_{i_m}^1, X_{j_1}^1, ...X_{j_s}^1, X_{i_{m+1}}^0, ..., X_{i_\ell}^0, \tag{79}$$

where $X_{j_1}^1, \ldots, X_{j_s}^1$ are the result of random sampling from $X_{i_1}^1, \ldots, X_{i_m}^1$ with replacement. Now, in order to estimate values $R(X_i), i = 1, ..., \ell$, we construct function $F_{\ell_{\mathrm{den}}}(x)$ from data (78) and function $F_{\ell_{\mathrm{num}}}(x)$ from data (79) and use the algorithm for density ratio estimation. For SVM method, in order to balance data, we have to maximize (73) subject to constraints (74) and (75).

8. Problem of Local Learning

In 1992, the following problem of local learning was formulated (Bottou and Vapnik, 1992): given data

$$(x_1, y_1), ..., (x_\ell, y_\ell) \tag{80}$$

generated according to an unknown density function

$$p_0(y, x) = p_0(y|x)p_0(x),$$

find the decision rule that minimizes risk *in a vicinity of the given point* x_0. Using some heuristic concept of vicinity of given points, the corresponding algorithm was developed. It was demonstrated (Bottou and Vapnik, 1992), (Vapnik and Bottou, 1993) that local learning is often more accurate than the global learning.

In this Section, we present a reasonable definition of the concept of locality, and we solve the problem of constructing local rules. Our goal is to use data (80) for constructing a rule that is accurate for vectors distributed according to some $p_{loc}(x)$, for example, according to the multidimensional normal distribution

$$p_{loc}(x) = N(x_0, \sigma I) = \frac{1}{(2\pi)^{m/2}\sigma^m} \prod_{k=1}^{m} \exp\left\{-\frac{(x^k - x_0^k)}{2\sigma^2}\right\},$$

where $x_0 = (x_0^1, ..., x_0^m)$ is the vector of means, $\sigma > 0$ and identity matrix I are known parameters of multi-dimensional normal distribution (they are specified by our concept of vicinity point x_0). We denote by $p_{loc}(y, x)$ the density function

$$p_{loc}(y, x) = p_0(y|x)N(x_0, \sigma I).$$

Therefore, the goal of local learning is to find, in the set of functions $f(x,\alpha), \alpha \in \Lambda$, the function $f(x,\alpha_n)$ that minimizes the functional

$$T_{loc}(\alpha) = \int (y - f(x,\alpha))^2 p_{loc}(y,x) dy dx$$

instead of the functional

$$T_0(\alpha) = \int (y - f(x,\alpha))^2 p_0(y,x) dy dx,$$

as it is formulated in classical (global) learning paradigm.

We rewrite functional T_{loc} as follows:

$$T_{loc}(\alpha) = \int (y - f(x,\alpha))^2 R(x) p_0(y,x) dy dx,$$

where we have denoted

$$R(x) = \frac{p_{loc}(x)}{p_0(x)}.$$

To minimize the functional T_{loc} given data obtained according to $p_0(y,x)$, we minimize the empirical risk

$$\hat{R}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i,\alpha))^2 R(x_i).$$

To define this functional explicitly, we have to estimate the ratio of two densities, one of which (the density that specifies vicinity of point x_0) is known, and another one is unknown but elements x of data obtained according to that unknown density $p_0(x)$ are available from the training set.

This problem of density ratio estimation, however, differs from the one considered in Section 6.3. It requires solving the integral equation when the right-hand side of equation is precisely defined, whereas the operator is defined approximately:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - X_i) R(x_i) \approx \int_{-\infty}^{x} N(x_0, \sigma I) dx.$$

The integral in the right-hand can be expressed as the product of m (where m is the dimensionality of space X) functions

$$\mathrm{Erf}(x_* | x_0, \sigma I) = \frac{2^m}{(\pi)^{m/2}} \prod_{k=1}^{m} \int_0^{x_*^k} \exp\left\{ -\frac{(x^k - x_0^k)^2}{2\sigma^2} dx^k \right\}.$$

Note that function $\mathrm{Erf}(x|x_0,\sigma I)$ can be easily computed using the so-called function $\mathrm{erf}(x)$:

$$\mathrm{erf}(x^k) = \frac{2}{\sqrt{\pi}} \int_0^{x^k} \exp\{-t^2\} dt,$$

which is tabulated.

Using the method of estimation of density ratio, in order to estimate vector $R = (R(x_1), ..., R(x_\ell))$, one has to minimize the functional

$$R^T V R - 2\ell R^T U + \gamma R^T K^+ R,$$

where we have denoted by $U = (U_1, ..., U_\ell)^T$ the vector with coordinates

$$U_i = \prod_{k=1}^{m} \sum_{t=1}^{\ell} \theta(x_t^k - x_i^k) \mathrm{Erf}(x_t^k | x_0, \sigma)$$

subject to constraints

$$R^T 1 = \ell$$

and constraints

$$R \geq 0_\ell.$$

In SVM technology, the V-matrix method requires to estimate values $R(X_i)$ in the points of observations first, and then to solve the SVM problem itself using the data adaptation technique described in Section 6.3.

9. Comparison with Classical Methods

In this paper, we introduced a new unified approach to solution of statistical inference problems based on their direct settings. We used rigorous mathematical techniques to solve them. Surprisingly, all these problems are amenable to relatively simple solutions.

One can see that elements of such solutions already exist in the basic classical statistical methods, for instance, in estimation of linear regression and in SVM pattern recognition problems.

9.1 Comparison with Linear Methods

Estimation of linear regression function is an important part of classical statistics. It is based on iid data

$$(y_1, X_1), ..., (y_\ell, X_\ell), \tag{81}$$

where y is distributed according to an unknown function $p(y|x)$. Distribution over vectors x is a subject of special discussions: it could be either defined by an unknown $p(x)$ or by known fixed vectors. It is required to estimate the linear regression function

$$y = w_0^T x.$$

Linear estimator. To estimate this function, classical statistics uses *ridge regression method* that minimizes the functional

$$R(w) = (Y - \mathbf{X}w)^T (Y - \mathbf{X}w) + \gamma(w, w), \tag{82}$$

where $\mathbf{X}$ is the $(\ell \times n)$-dimensional matrix of observed vectors X, and Y is the $(\ell \times 1)$-dimensional matrix of observations y. This approach also covers the least squares method (for which $\gamma = 0$).

When observed vectors X in $\mathbf{X}$ are distributed according to an unknown $p(x)$, method (81) is consistent under very general conditions.

The minimum of this functional has the form

$$w_\ell = (\mathbf{X}^T\mathbf{X} + \gamma I)^{-1}\mathbf{X}^T Y. \tag{83}$$

However, estimate (82) is not necessarily the best possible one.

The main theorem of linear regression theory, the *Gauss-Markov* theorem, assumes that input vectors X in $\mathbf{X}$ (81) are fixed (they are not random!). Below we formulate it in a slightly more general form.

Theorem. *Suppose that the random values $(y_i - w_0^T X_i)$ and $(y_j - w_0^T X_j)$ are uncorrelated and that the bias of estimate (82)*

$$\mu = E_y(w_\ell - w_0).$$

Then, among all linear[8] *estimates with bias*[9] *μ, estimate (82) has the smallest expectation of squared deviation:*

$$E_y(w_0 - w_\ell)^2 \leq E_y(w_0 - w)^2, \quad \forall w.$$

Generalized linear estimator. Gauss-Markov model can be extended in the following way. Let ℓ-dimensional vector of observations Y be defined by fixed vectors X and additive random noise $\Omega = (\varepsilon_1, ..., \varepsilon_\ell)^T$ so that

$$Y = \mathbf{X}w_0 + \Omega,$$

where the noise vector $\Omega = (\varepsilon_1, ..., \varepsilon_\ell)^T$ is such that

$$E\Omega = 0, \tag{84}$$

$$E\Omega\Omega^T = \Sigma. \tag{85}$$

Here, the noise values at the different points X_i and X_j of matrix $\mathbf{X}$ are correlated and the correlation matrix Σ is *known* (in the classical Gauss-Markov model, it is identity matrix $\Sigma = I$). Then, instead of estimator (82) minimizing functional (81), one minimizes the functional

$$R(w) = (Y - \mathbf{X}w)^T\Sigma^{-1}(Y - \mathbf{X}w) + \gamma(w, w). \tag{86}$$

This functional is obtained as the result of de-correlation of noise in (83), (84). The minimum of (85) has the form

$$\hat{w}_* = (\mathbf{X}^T\Sigma^{-1}\mathbf{X} + \gamma I)^{-1}\mathbf{X}^T\Sigma^{-1}Y. \tag{87}$$

This estimator of parameters w is an improvement of (82) for correlated noise vector.

***V*-matrix estimator of linear functions.** The method of solving regression estimation problem (ignoring constraints) with V matrix leads to the estimate

$$\hat{w}_{**} = (\mathbf{X}^T V\mathbf{X} + \gamma I)^{-1}\mathbf{X}^T VY.$$

[8] Note that estimate (83) is linear only if matrix X is fixed.

[9] Note that when $\gamma = 0$ in (82), the estimator (82) with $\gamma = 0$ is unbiased.

The structure of the V-matrix based estimate is the same as those of linear regression estimates (82) and (86), except that the V-matrix replaces identity matrix in (82) and inverse covariance matrix in (86).

The significant difference, however, is that both classical models were developed for the known (fixed) vectors X, while V-matrix is defined for random vectors X and is computed using these vectors. It takes into account the information that classical methods ignore: the domain of regression function and the geometry of observed data points. The complete solution also takes into accounts the constraints that reflects the belief in estimated prior knowledge about the solution.

9.2 Comparison with L_2-SVM (Non-Linear) Methods

For simplicity, we discuss in this section only pattern recognition problem; we can use the same approach for the non-linear regression estimation problem.

The pattern recognition problem can be viewed as a special case of the problem of conditional probability estimation. Using an estimate of conditional probability $p(y=1|x)$, one can easily obtain the classification rule

$$f(x) = \theta(p(y=1|x) - 1/2).$$

We now compare the solution $f(x)$ with

$$f(x) = A^T \mathcal{K}(x)$$

obtained for conditional probability problem with the same form of solution that defines SVM.

The coefficients A for L_2-SVM have the form (Saunders et al., 1998), (Suykens and Vandewalle, 1999)

$$A = (K + \gamma I)^{-1} Y. \tag{88}$$

If V-matrix method ignores the prior knowledge about the properties of conditional probability function, the coefficients of expansion have the form

$$A = (KV + \gamma I)^{-1} VY. \tag{89}$$

It is easy, however, to incorporate the existing constraints into both solutions.

In order to find the standard hinge-loss SVM solution (Vapnik, 1995), we have to minimize the quadratic form

$$-A^T \mathcal{Y} K \mathcal{Y} A + 2A^T \mathbf{1}_\ell$$

with respect to A subject to the box constraint

$$\mathbf{0}_\ell \leq A \leq C\mathbf{1}_\ell$$

and the equality constraint

$$A^T \mathcal{Y} \mathbf{1}_\ell = 0,$$

where C is the (penalty) parameter of the algorithm, and $\mathcal{Y}$ is $(\ell \times \ell)$-dimensional diagonal matrix with $y_i \in \{-1, +1\}$ from training data on its diagonal (see formulas (71), (72) , (73), (74), and (75) with $R(x_i) = 1$ in (71) and (75)).

In order to find the values of conditional probability, we also have to minimize the quadratic form

$$\Phi^T(V + \gamma K^+)\Phi - 2\Phi VY,$$

with respect to Φ subject to the box constraints[10]

$$\mathbf{0}_\ell \leq \Phi \leq \mathbf{1}_\ell$$

and the equality constraint

$$\Phi^T \mathbf{1}_\ell = \ell p_\ell,$$

where $\gamma > 0$ is the (regularization) parameter of the algorithm in the objective function (as C for SVM).

The essential difference between SVM and V-matrix method is that the constraints in SVM method appear due to necessary technicalities (related to Lagrange multiplier method[11]) while in V-matrix method they appear as a result of incorporating existing prior knowledge about the solution: the classical setting of pattern recognition problem does not include such prior knowledge[12].

The discussion above indicates that, on one hand, the computational complexity of estimation of conditional probability is not higher than that of standard SVM classification, while, on the other hand, the V-estimate of conditional probability takes into account not only the information about the geometry of training data (incorporated in V-matrix) but also the existing prior knowledge about solution (incorporated in the constraints above).

This leads to the following question:

Can V-matrix method replace SVM method for pattern recognition?

The answer to this question is not obvious. In the mid-1990s, the following Imperative was formulated (Vapnik, 1995), (Vapnik, 1998):

> *While solving problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you need, but not a more general one. It is quite possible that you have enough information to solve a particular problem of interest well, but not enough information to solve a general problem.*

[10] Often one has stronger constraints

$$\mathbf{a}_\ell \leq \Phi \leq \mathbf{b}_\ell,$$

where $\mathbf{0}_\ell \leq \mathbf{a}_\ell$ and $\mathbf{b}_\ell \leq \mathbf{1}_\ell$ are given (by experts) as additional prior information.

[11] The Lagrange multiplier method was developed to find the solution in the *dual optimization space* and constraints in SVM method are related to Lagrange multipliers. Computationally, it is much easier to obtain the solution in the dual space given by (73), (74), (75) than in the *primal space* given by (71), (72). As shown by comparisons (Osuna and Girosi, 1999) of SVM solutions in primal and dual settings, (1) solution in primal space is more difficult computationally, (2) the obtained accuracies in both primal and dual spaces are about the same, (3) the primal space solution uses significantly fewer support vectors, and (4) the large number of support vectors in dual space solution is caused by the need to maintain the constraints for Lagrange multipliers.

[12] The only information in SVM about the solution are the constraints $y_i f(x_i, \alpha) \geq 1 - \xi_i$, where $\xi_i \geq 0$ are (unknown) slack variables (Vapnik and Izmailov, 2015). However, this information does not contain any prior knowledge about the function f.

Solving (ill-posed) conditional probability problem instead of pattern recognition problem might appear to contradict this Imperative. However, while estimating conditional probability, one uses prior knowledge about the solution, and applies rigorous approaches, whereas the SVM setting does not take that knowledge into account and is based, instead, on justified heuristic approach of large margin. Since these two approaches leverage different factors and thus cannot be compared theoretically, it is important to compare them empirically.

9.3 Experimental Comparison of I-Matrix (L_2 SVM) and V-matrix Methods

In this section, we compare the L_2-SVM based method with V-matrix based method for estimation of one-dimensional conditional probability functions. Let the data be generated by an unknown probability density function $p(x,y)=p(y|x)p(x)$, where $x\in X,\ y\in\{0,1\}$. Then the regression function $f_0(x)$ coincides with the conditional probability function $p(y=1|x)$, so the problem of estimating the conditional probability in the set $\{f(x,\alpha)\},\alpha\in\Lambda$ is equivalent to the problem of estimating the regression function on the data

$$(x_1,y_1),...,(x_\ell,y_\ell).$$

We use L_2-SVM method for the estimation of the non-linear regression function in the set $\{f(x,\alpha)\},\alpha\in\Lambda$ belonging to RKHS.

According to this method, in order to estimate the regression in the set of RKHS associated with the kernel $K(x_i,x)$, one has to find the parameters α_i of the function

$$f(x,\alpha)=\sum_{i=1}^{\ell}\alpha_i K(x_i,x)+\alpha_0$$

that minimize the functional

$$(Y-K\Lambda-\alpha_0 1)^T(Y-K\Lambda-\alpha_0 1)+\gamma\Lambda^T K\Lambda, \tag{90}$$

where we have denoted $Y=(y_1,...,y_\ell)^T$, $\Lambda=(\alpha_1,...,\alpha_\ell)^T$, by K is the matrix of elements $K(x_i,x_j),\ i,j=1,...,\ell$ and 1 is the ℓ-dimensional vector of ones.

Additionally, we take into account that (since regression coincides with conditional probability) the desired function satisfies $(\ell+1)$ constraints: one constraint of equality type

$$\frac{1}{\ell}\sum_{i,j=1}^{\ell}\alpha_i K(x_i,x_j)+\alpha_0=\frac{1}{\ell}\sum_{i=1}^{\ell}y_i, \tag{91}$$

and ℓ constraints of inequality type

$$0\le\sum_{i=1}^{\ell}\alpha_i K(x_i,x_j)+\alpha_0\le 1,\quad j=1,...,\ell, \tag{92}$$

forming L_2-SVM based method of conditional probability estimation.

The V-matrix based method of conditional probability estimation minimizes the functional

$$(Y-K\Lambda-\alpha_0 1)^T V(Y-K\Lambda-\alpha_0 1)+\gamma\Lambda^T K\Lambda \tag{93}$$

subject to the same constraints.

Therefore, the L_2-SVM method differs from V-matrix method by using identity matrix I instead of V matrix. Further, we call these method as I-matrix and V-matrix methods[13] In this Section, we present results of experimental comparisons of I-matrix and V-matrix methods. In our comparison, we consider two (one-dimensional) examples: estimating monotonic[14] and non-monotonic functions. In our experiments, we use the same kernel, namely, INK-spline of order 0:

$$K(x_i, x_j) = \min(x, x_i).$$

We can apply three versions of the solution for this problem:

1. Solutions that are defined by closed forms (ignoring the prior knowledge about the problem). These solutions are fast to obtain, without any significant computational problems.
2. Solutions that minimize the corresponding functionals while taking into account only the constraint of equality type.

   These solutions are also fast, without any significant computational problems. In this case one has to minimize functionals (90) and (93) choosing such α_0 for which equality constraints (91) holds true (Kuhn-Tucker condition)
3. Solutions that minimize functionals (90), (93) subject to all $\ell + 1$ constraints.

   These solutions require applying a full-scale quadratic optimization procedure. For large values of ℓ, it is not as simple computationally as previous two versions.

For our examples, all three solutions gave reasonably close results. Below we only report the results of the last one, the QP-solution.

Our first goal was to estimate the effect of using V-matrix (and compare it to I-matrix). To do this, we had to exclude the influence of the choice of regularization parameter γ. We did this by using two one-dimensional problems of estimating conditional probability functions: (1) monotonic function (Figure 1) and (2) non-monotonic one (Figure 2). For each problem, we generated 10,000 test examples and selected the best the possible (for the given training set) value of parameter γ. Figure 1 and Figure 2 present the result of approximation of conditional probability function for training sets of different sizes (48, 96,

[13] If one ignores the constraints, both methods (I-matrix method the V-matrix method) have closed form solutions. The solutions are (for I-matrix method $V = I$)

$$A = \left(WK + \frac{\gamma}{2} I\right)^{-1} WY,$$

where

$$W = V - c^{-1}(V\mathbf{1})(\mathbf{1}^T V), \qquad c = \mathbf{1}^T V \mathbf{1}.$$
$$\alpha_0 = c^{-1}\mathbf{1}^T V \left(Y - KA\right).$$

[14] Estimation of monotonic conditional probability function is important for pattern recognition problem since the VC dimension of the set of monotonically increasing (decreasing) functions equal to one independently of dimensionality.

192, 384) using the best γ for I-matrix method (left column) and V-matrix method (right column). In the figures, blue color corresponds to the true condition probability function, while black color corresponds to its approximations; red and green points in the horizontal axis correspond to two classes of the training set. In the Figures, we also show deviations of the approximations from the true conditional probability functions in both $L_1(\mu)$ and $L_2(\mu)$ metrics. In all our experiments we used the equal number of representatives of both classes.

These comparisons show that in all cases V-matrix method delivers better solution.

Subsequently, we compared V-matrix and I-matrix methods when the parameter γ is selected using the cross-validation technique on training data (6-fold cross validation based on maximum likelihood criterion): Figure 3 and Figure 4. Here also V-matrix method performs better than I-matrix method. The more training data is used, the larger is the advantage of the V-matrix method.

It is especially important that, in all our experiments, V-matrix method produced more smooth approximations to the true function than I-matrix method did. This is due to incorporation of the geometry of the training data into the solution.

Acknowledgments

This material is based upon work partially supported by AFRL and DARPA under contract FA8750-14-C-0008. Any opinions, findings and / or conclusions in this material are those of the authors and do not necessarily reflect the views of AFRL and DARPA.

We thank Professor Cherkassky, Professor Gammerman, and Professor Vovk for their helpful comments on this paper.

Appendix A. Appendix: V-Matrix for Statistical Inference

In this section, we describe some details of statistical inference algorithms using V-matrix. First, consider algorithms for conditional probability function $P(y|x)$ estimation and regression function $f(x)$ estimation given iid data

$$(y_1, X_1), ..., (y_\ell, X_\ell) \tag{94}$$

generated according to $p(x, y) = p(y|x)p(x)$. In (94), $y \in \{0, 1\}$ for the problem of conditional probability estimation, and $y \in R^1$ for the problems of regression estimation and density ratio estimation. Our V-matrix algorithm consists of the following simple steps.

A.1 Algorithms for Conditional Probability and Regression Estimation

Step 1. Find the domain of function. Consider vectors

$$X_1, ..., X_\ell \tag{95}$$

from training data. By a linear transformation in space $\mathcal{X}$, this data can be embedded into the smallest rectangular box with its edges parallel to coordinate axes. Without loss of generality, we also chose the origin of coordinate y such that all $y_i \in [0, \infty]$, $i = 1, ..., \ell$ are non-negative.

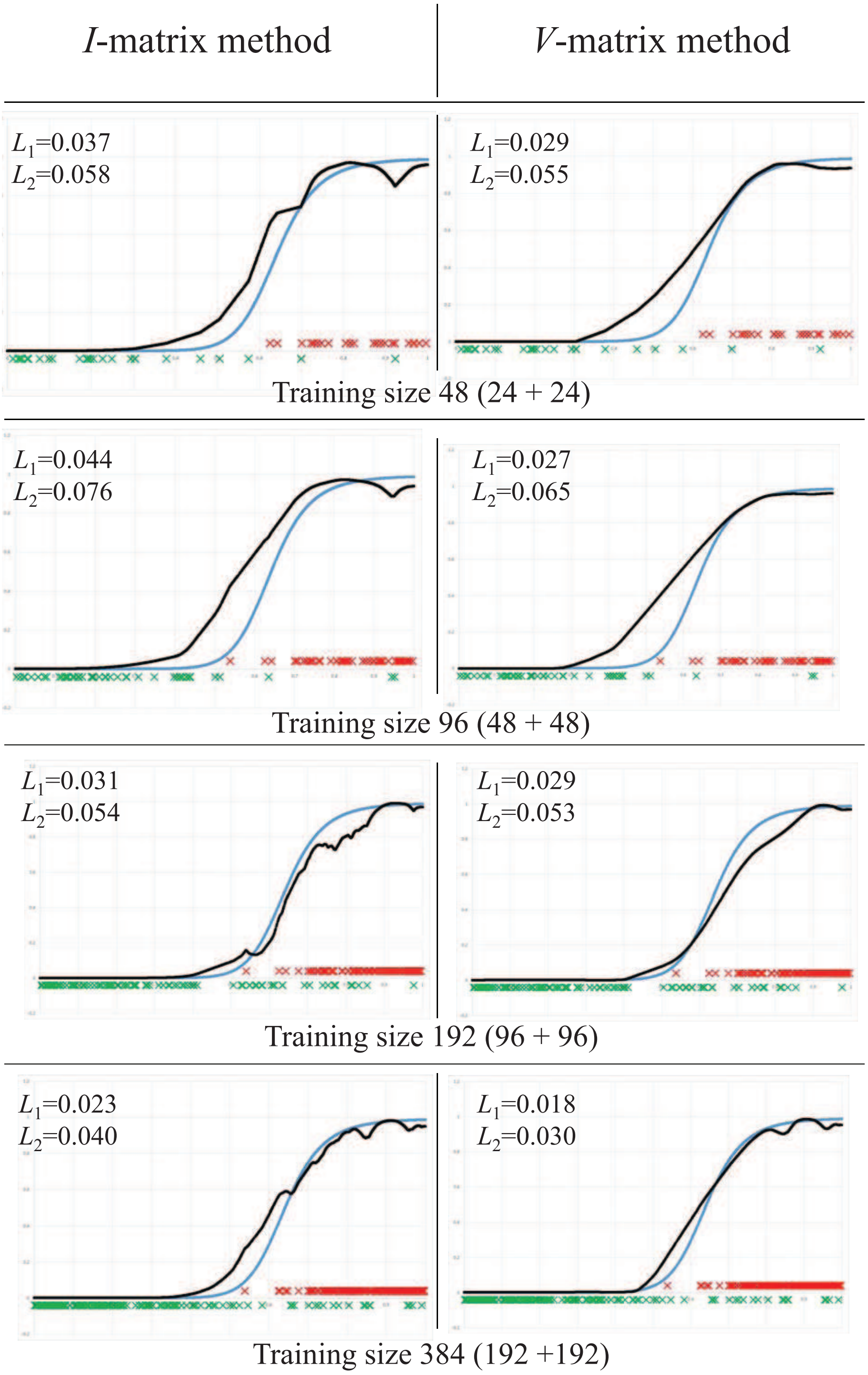

Figure 1: Comparison of I-matrix and V-matrix methods where regularization parameters γ were selected on validation set of size 10,000.

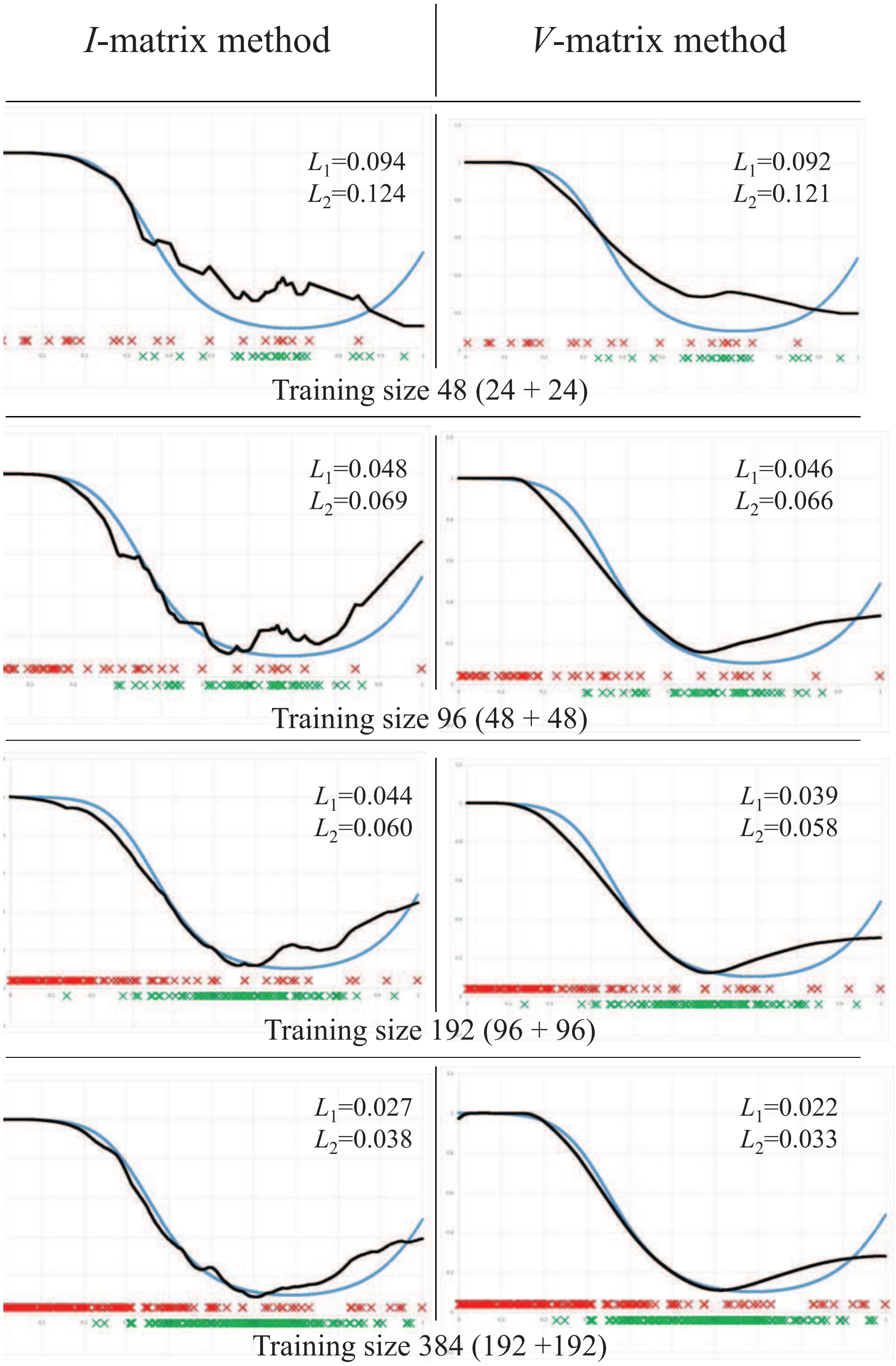

Figure 2: Comparison of I-matrix and V-matrix methods where regularization parameters γ were selected on validation set of size 10,000.

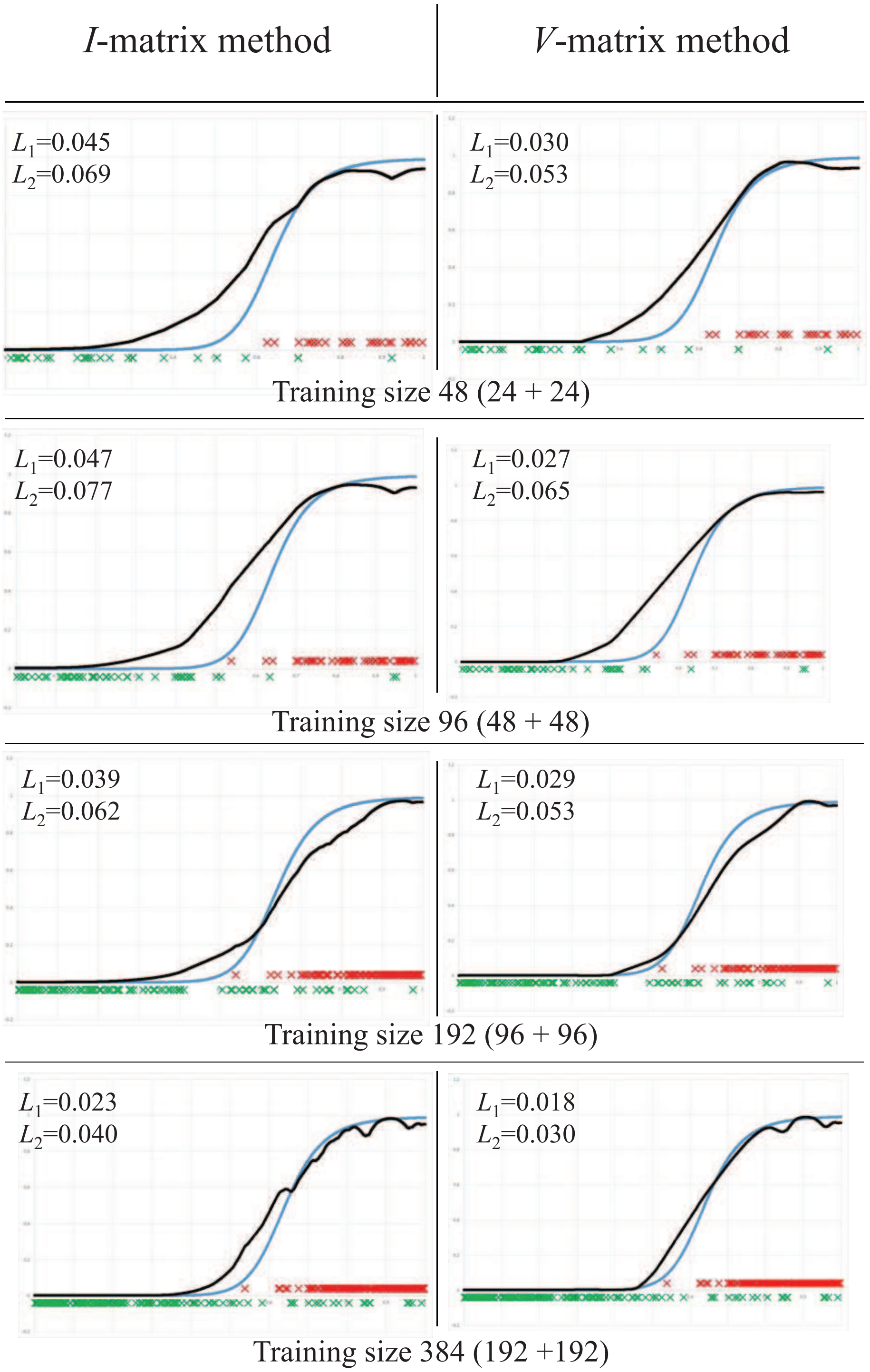

Figure 3: Comparison of I-matrix and V-matrix methods where regularization parameters γ were selected by cross-validation on training set.

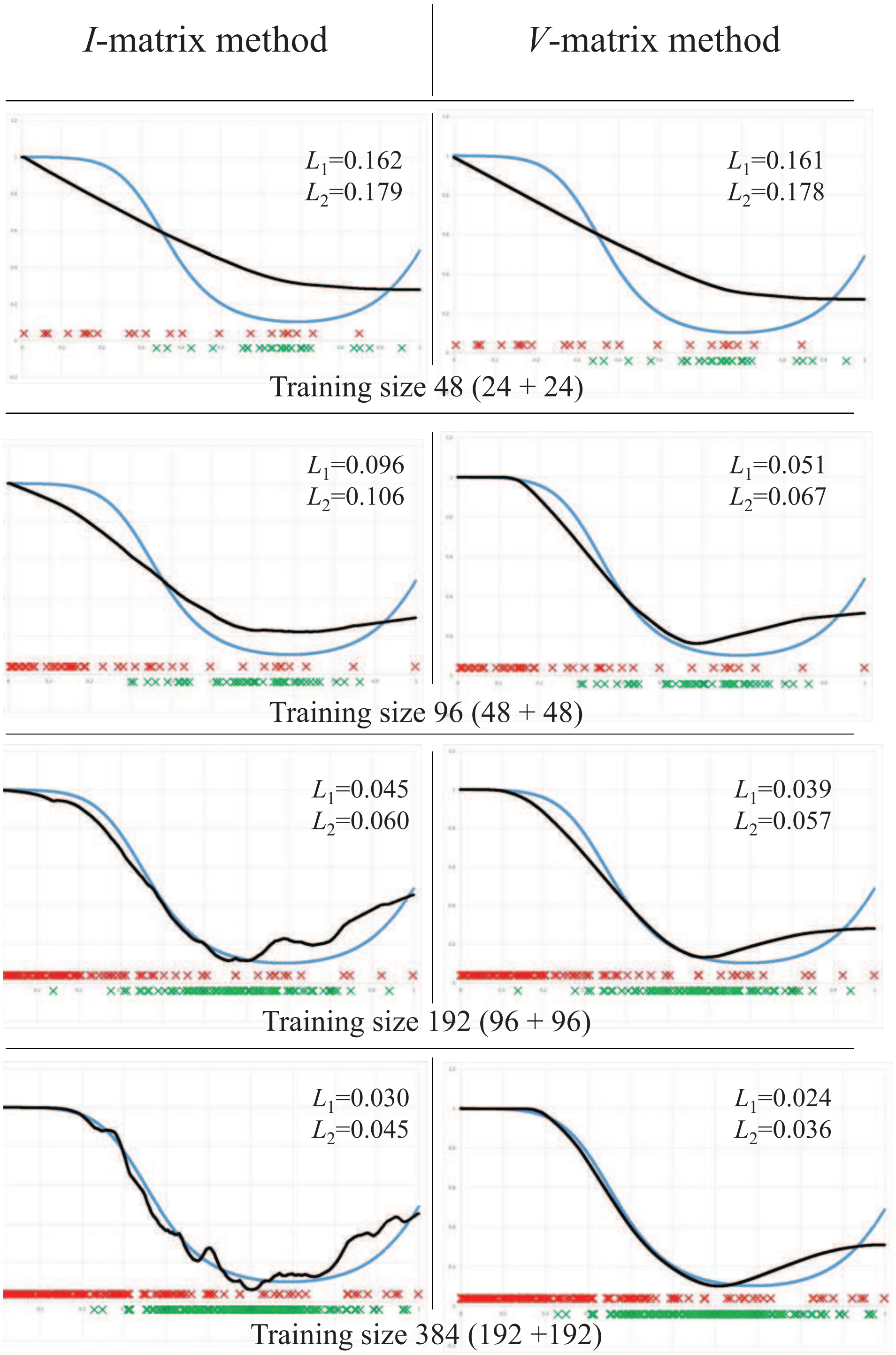

Figure 4: Comparison of I-matrix and V-matrix methods where regularization parameters γ were selected by cross-validation on training set.

Further we assume that data (95) had been preprocessed in this way.

Step 2. Find the functions $\mu(x^k)$. Using preprocessed data (95), construct for any coordinate x^k of the vector x the piecewise constant function

$$\mu_k(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x^k - X_i^k).$$

Step 3. Find functions $\sigma(x^k)$. For any coordinate of $k = 1, ..., d$ find the following:

1. The value

$$\hat{y}_{av} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i$$

(for pattern recognition problem, $\hat{y}_{av} = p_\ell$ is the fraction of training samples from class $y = 1$).

2. The piecewise constant function

$$F_*(x^k) = \frac{1}{\ell\hat{y}_{av}}\sum_{i=1}^{\ell} y_i\theta(x - X_i)$$

(For pattern recognition problem, function $F_*(x^k) = P(x^k|y = 1)$ estimates cumulative distribution function of x^k for samples from class $y = 1$).

3. The piecewise constant function

$$\sigma^k(x) = \Big(F_*(x^k)(1 - F_*(x^k)) + \varepsilon\Big)^{-1}.$$

Step 4. Find elements of V-matrix. Calculate the values

$$V_{ij}^k = \int \theta(x^k - X_i^k)\theta(x^k - X_j^k)\sigma(x^k)d\mu(x^k) = \int_{\max\{X_i^k, X_j^k\}}^{\infty} \sigma(x^k)d\mu(x^k).$$

Since both $\sigma(x^k)$ and $\mu(x^k)$ are piecewise constant functions, the last integral is a sum of constants.

Step 5. Find V-matrix. Compute elements of V-matrix as

$$V_{ij} = \prod_{k=1}^{d} V_{ij}^k.$$

Remark 1. Since V-matrix in the problems of conditional probability and regression estimation is scale-invariant, one can multiply all elements of this matrix by a fixed constant in order to keep the values of matrix elements within reasonable bounds for subsequent computations.

Remark 2. Any diagonal element V_{tt}^k is not less than elements of the corresponding row V_{tj}^k and column V_{jt}^k. Therefore, in order to compute V-matrix in multi-dimensional case, it is reasonable to compute its diagonal elements first and, if they are small, just to replace the entries in the corresponding row and column with zeros.

It is possible (especially for large d) that V-matrix can have dominating diagonal elements. In this case, V-matrix can be approximated by a diagonal matrix. This is equivalent to the weighted least square method where weights are defined by the diagonal values V_{tt}.

Step 6. Find the values of conditional probability or the values of regression at the points of observation. Solve the quadratic optimization problem defined in the corresponding sections (in Section 6.4).

Step 7. Find the conditional probability or regression function. Solve interpolation problem defined in Section 6.4.

A.2 Algorithms for Density Ratio Estimation

For the problem of density ratio estimation, the algorithm requires the following modifications:

Step 1a. Find the domain of function. Domain of function is defined using data

$$X_1, ..., X_{\ell_{\text{den}}}, X_1', ..., X_{\ell_{\text{num}}}', \tag{96}$$

where training vectors X_i and X_j' are distributed according to $F_{\text{den}}(x)$ and $F_{\text{num}}(x')$, respectively.

Step 2a. Find the functions $\mu(x^k)$. Using (preprocessed) data (96), construct for coordinate $x^k,\ k = 1, ..., d$ of vector x the piecewise constant function

$$\mu_k(x) = \frac{1}{(\ell_{\text{den}} + \ell_{\text{num}})} \left(\sum_{i=1}^{\ell_{\text{den}}} \theta(x^k - X_i^k) + \sum_{i=1}^{\ell_{\text{num}}} \theta(x^k - X_i'^k) \right).$$

Step 3a. Find functions $\sigma(x^k)$. For any coordinate $x^k,\ k = 1, ..., d$ find:

- the piecewise constant function

$$F_{**}(x^k) = \frac{1}{\ell_{\text{num}}} \sum_{j=1}^{\ell_{\text{num}}} \theta(x - X_j');$$

- the piecewise constant function

$$\sigma(x^k) = \Big(F_{**}(x^k)(1 - F_{**}(x^k)) + \varepsilon \Big)^{-1},$$

where $\varepsilon > 0$ is a small value.

Step 4a. Find the V-matrix and V^*-matrix. Estimate the matrices using expressions from corresponding sections.

Step 5a. Find the values of density ratio function at the points of observation. Solve the quadratic optimization problem defined in corresponding sections.

Step 6a. Find the density ratio function. Solve the interpolation problem defined in Section 6.4 (if estimated values of density ratio in ℓ_{den} points are not sufficient for the application, and the function itself has to be estimated).

A.3 Choice of Regularization Parameter

The value of regularization parameter γ can be selected using standard cross-validation techniques.

For conditional probability estimation, one can look for maximization of likelihood rather than for minimization of error rate. This leads to a more accurate estimate of conditional probability function.

References

L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

G. Brown, A. Pocock, M. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.*, 13: 27–66, January 2012.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

M. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, pages 823–830, 2012.

Y. Kawahara and M. Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In H. Park, S. Parthasarathy, H. Liu, and Z. Obradovic, editors, *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)*, pages 389–400, Sparks, Nevada, USA, Apr. 30–May 2 2009.

G. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, pages 495–502, 1970.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

E. Osuna and F. Girosi. Reducing the run-time complexity in support vector machines. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods*, pages 271–283. MIT Press, Cambridge, MA, USA, 1999.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426, London, UK, UK, 2001. Springer-Verlag.

A. Stefanyuk. Estimation of the likelihood ratio function in the “disorder” problem of random processes. *Automation and Remote Control*, (9):53–59, 1986.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning.* Cambridge University Press, Cambridge, UK, 2012.

J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9(3):293–300, June 1999.

A. Tikhonov and V. Arsenin. *Solutions of Ill-Posed Problems.* W.H. Winston, 1977.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

V. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

V. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition (in Russian).* Nauka, Moscow, 1974.

V. Vapnik and R. Izmailov. Learning with intelligent teacher: Similarity control and knowledge transfer. In A. Gammerman, V. Vovk, and H. Papadopoulos, editors, *Statistical Learning and Data Sciences*, volume 9047 of *Lecture Notes in Computer Science*, pages 3–32. Springer International Publishing, 2015.