Monte Carlo Methods Appl. Vol. 15 No. 2 (2009), pp. 135–144
DOI 10.1515 / MCMA.2009.007

On importance sampling in the problem of global optimization

Trifon I. Missov and Sergey M. Ermakov

Abstract. Importance sampling is a standard variance reduction tool in Monte Carlo integral evaluation. It postulates estimating the integrand just in the areas where it takes big values. It turns out this idea can be also applied to multivariate optimization problems if the objective function is non-negative. We can normalize it to a density function, and if we are able to simulate the resulting p.d.f., we can assess the maximum of the objective function from the respective sample.

Keywords. Global optimization, importance sampling, Δ^2-distribution, D-optimal designs.

AMS classification. 65C05, 62K05, 68U20.

1. Introduction

Global optimization is usually performed by using lattice-based search methods (see, for example, [15] and [12]). However, they are applicable just to functions with unique global extremum and are feasible just in a bounded search region. Their most general framework incorporates at each step the simulation of a sequence of uniformly distributed random vectors, in which the objective function is evaluated. Then, in a vicinity of the record value a new procedure of the same type is executed. In addition, the use of low discrepancy deterministic vector sequences leads to more accurate results. Nevertheless, such methods converge slowly and, for example, Zhigljavsky (1987) provides an estimate of the minimum number of algorithm steps N for reaching an ε-vicinity of the global optimum x^* with a $1-\gamma$, $0<\gamma<1$ level of certainty:

$$N(\varepsilon,\gamma)=\left\lceil \frac{\ln\gamma}{\ln\left(1-\mu(B(x^*,\varepsilon))\right)} \right\rceil, \tag{1.1}$$

where $B(x^*,\varepsilon)$ is the s-dimensional ball of radius ε centered at the global optimum.

In the case when we have information about the approximate location of the optimum, we can simulate points out of some other (non-uniform) distribution. For instance, we can choose a p.d.f., which incorporates such information. In particular, if the objective function is non-negative and, consequently, we search for its maximum, we can sample from a p.d.f., resulting from the normalization of the objective function

itself. Hence, the mode of the resulting distribution will capture the maximum point we are searching for.

Let us illustrate this idea on a simple example. Suppose we search for the maximum of the function $f(x) = x^n$ in $[0, 1]$. First, we construct a p.d.f. $p(x) = (n + 1)x^n$, whose mode is apparently $x = 1$ (the corresponding maximum equal to $n + 1$). If we optimize $f(x)$ by evaluating it in uniformly distributed points in $[0, 1]$, we can get into an ε-vicinity of 1 after $\asymp 1/\varepsilon$ steps. On the other hand, if we choose the points of $f(x)$ evaluation from $p(x)$, we will need just $\approx 1/((n + 1)\varepsilon)$ steps. As a result, the bigger n, i.e. the maximum of the function we want to evaluate, the smaller the number of algorithm steps. This simple idea addresses a class of problems in which particular p.d.f. simulation can lead to efficient optimization algorithms. More specifically, it suits the maximization of non-negative functions with distinct and high maximum values.

2. The problem of exact D-optimization

Let us focus on a particular class of problems in statistics in which we can apply the importance sampling idea. Assume a set X with a σ-finite measure μ, $x_1, \ldots, x_n \in X$, and an orthonormalized system of functions $\varphi_1, \ldots, \varphi_m$ in $L^2(X, \mu)$, $n > m$. We are interested in the maximum of the following function:

$$f_{n,m}(Q) = \det \left\| \sum_{i=1}^{n} \varphi_k(x_i)\varphi_l(x_i) \right\|_{k,l=1}^{m}, \tag{2.1}$$

where $Q = (x_1, \ldots, x_n)$.

This determinant, known as the information design matrix, plays an important role in the design of experiment problems. In particular, the argument of its maximum corresponds to the D-optimal design. In regression analysis a D-optimal design provides a vector parameter estimate with the lowest volume of its corresponding variance ellipsoid. Solving D-optimization problems is a difficult task, which is most often approached by studying continuous designs and applying the equivalence theory of Kiefer and Wolfowitz ([7], 1959). As a result, the D-optimal design is determined by rounding its corresponding continuous one. However, exact D-optimal designs are known just in a several one-dimensional cases like, for instance, polynomial and trigonometric regression. Moreover, continuous design theory does not cover the $m = n$ case. For an arbitrary region, Wynn ([14], 1970) and Fedorov ([6], 1971) proposed an iterative numeric procedure, which is, though, associated with highly operation consuming matrix transforms. As a result, it is important to develop approaches that evade such kind of complexity. The procedure we would like to offer in this paper might serve at least as an initial step for allocating the maximum of the information design matrix. After that a number of local optimization gradient methods could be applied for specifying the maximum more accurately.

Simulating distributions with densities proportional to (2.1) is a difficult task. However, Bogues et al. ([1], 1981), Ermakov and Missov ([3], 2005), and Missov ([8], 2007) proposed methods evading the computation of determinants of high orders. Let us consider the following density function

$$\Delta^2_{n,m}(Q) = \frac{(n-m)!}{n!} \det \left\| \sum_{i=1}^{n} \varphi_k(x_i)\varphi_l(x_i) \right\|^m_{k,l=1}. \tag{2.2}$$

It suits our framework as the maximum of the determinant in (2.2) increases rapidly with m and n. As a result, by applying the above stated idea, we can find at least a good initial approximation of the objective function's global maximum.

Note that (2.2) is a generalization of the Δ^2 (or Ermakov–Zolotukhin, [5]) distribution in the case when the number n of points $x_1, \ldots, x_n$ exceeds the number m of functions $\varphi_1, \ldots, \varphi_m$. The original Ermakov–Zolotukhin distribution is designed for $n = m$. Coherently, we shall further call (2.2) the p.d.f. of the generalized Δ^2-distribution.

3. Simulation of the generalized Δ^2-distribution

The simulation of the generalized Δ^2 distribution is based on the algorithm of Ermakov and Missov ([3], 2005). First, we orthonormalize $\varphi_1, \ldots, \varphi_m$ and keep the same notation for the resulting system. Then we subsequently simulate the conditional densities of the generalized Δ^2-distribution. Each conditional density is a composition of distributions with iteratively computable coefficients.

When the region X has a complex structure, orthonormalizing $\varphi_1, \ldots, \varphi_m$ might not be trivial. A possible solution in this case is to inscribe X within another region Y in which this procedure can be easily performed. Afterwards, we simulate points in Y narrowing the distribution to X. Note that affine transformations of X result in affine transformations of $\varphi_1, \ldots, \varphi_m$. As a result, by finding the optimum in X, we know the solution for any other region $Y = F(X)$, where F is any affine transformation.

Depending on the orthonormalized system of functions $\varphi_1, \ldots, \varphi_m$ in $L^2(X, \mu)$, (2.1) can have multiple maxima. This is the case especially when $n \neq km$, $k \in \mathbb{Z}$. We will not focus on this problem analytically, but rather show numerical examples unveiling some basic mechanisms.

Last but not least, sampling from the generalized Δ^2-distribution leads to the following minimum number of algorithm steps $N = N(\varepsilon, \gamma)$ for reaching an ε-vicinity of the global maximum with a fixed level of certainty γ:

$$N(\varepsilon, \gamma) = \left\lceil \frac{\ln \gamma}{\ln \left(1 - \max\limits_{Q} \Delta^2_{n,m}(Q) \cdot \mu(B(x^*, \varepsilon)) \right)} \right\rceil. \tag{3.1}$$

As the maximum of $\Delta^2_{n,m}(Q)$ increases rapidly with m and n, the number of algorithm steps is considerably smaller than the one in (1.1). The latter is illustrated in Tables 1, 2, and 3, which provide comparison for the N values in pure random and generalized Δ^2 search for $s = 1$, $n = m = 1, \ldots, 11$, and $\varphi_1, \ldots, \varphi_m$ – the orthonormalized system of Legendre polynomials. This is exactly the case when $\Delta^2_{n,m}(Q)$ is equal to the square of the Vandermonde determinant, adjusted for normalizing terms. The maximum of the squared Vandermonde determinant is well known, [13]:

$$\frac{2^2 \cdot 3^3 \ldots n^n \cdot 2^2 \cdot 3^3 \cdot (n-2)^{n-2}}{3^3 \cdot 5^5 \ldots (2n-3)^{2n-3}}.$$

4. Numerical examples

Let us first illustrate the important sampling idea in the well-known one-dimensional case, when $X = [0, 1]$, μ is the Lebesgue measure, $m = 3$, and $\varphi_1, \ldots, \varphi_m$ are orthonormalized univariate polynomials, i.e.

$$\varphi_1(x) = 1, \qquad \varphi_2(x) = \sqrt{3}\,(2x - 1), \qquad \varphi_3(x) = \sqrt{5}\,(6x^2 - 6x + 1).$$

In these settings, when $n = m = 3$, (2.1) reaches its maximum for $x_1 = 0$, $x_2 = 1/2$, $x_3 = 1$ ([4]). In general, if $n = km$, $k \in \mathbb{Z}$, maximizing (1) is equivalent to choosing k times each of these m points. Suppose $n \equiv 1 (\mathrm{mod}\, m)$, for instance, $n = 4$. Simulating samples of size $N = \{500, 1000, 2000, 5000, 10000\}$ from the respective generalized

n	$\varepsilon = 0.05$	$\varepsilon = 0.025$	$\varepsilon = 0.01$	$\varepsilon = 0.001$
2	94	380	2382	238391
3	634	5084	79463	$7.95 \cdot 10^7$
4	5268	84313	$3.29 \cdot 10^6$	$3.29 \cdot 10^{10}$
5	50383	$1.61 \cdot 10^6$	$1.57 \cdot 10^8$	$1.57 \cdot 10^{13}$
6	536860	$3.44 \cdot 10^7$	$8.39 \cdot 10^9$	$8.99 \cdot 10^{15}$
7	$6.25 \cdot 10^6$	$8.00 \cdot 10^8$	$4.88 \cdot 10^{11}$	$6.07 \cdot 10^{18}$
8	$7.83 \cdot 10^7$	$2.00 \cdot 10^{10}$	$3.06 \cdot 10^{13}$	$4.70 \cdot 10^{21}$
9	$1.05 \cdot 10^9$	$5.36 \cdot 10^{11}$	$2.08 \cdot 10^{15}$	$4.12 \cdot 10^{24}$
10	$1.48 \cdot 10^{10}$	$1.52 \cdot 10^{13}$	$1.52 \cdot 10^{17}$	$4.01 \cdot 10^{27}$
11	$2.20 \cdot 10^{11}$	$4.50 \cdot 10^{14}$	$1.18 \cdot 10^{19}$	$4.31 \cdot 10^{30}$

Table 1. Sampling from the generalized Δ^2-distribution. Minimum number of algorithm steps N for reaching an ε-vicinity of the D-optimal design with a level of certainty $1 - \gamma = 0.95$ for $X = [-1, 1]$, $m = n$.

n	$\varepsilon = 0.05$	$\varepsilon = 0.025$	$\varepsilon = 0.01$	$\varepsilon = 0.001$
2	380	1524	9534	953570
3	5720	45770	715177	$7.15 \cdot 10^8$
4	97128	$1.55 \cdot 10^6$	$6.07 \cdot 10^7$	$6.07 \cdot 10^{11}$
5	$1.82 \cdot 10^6$	$5.83 \cdot 10^7$	$5.69 \cdot 10^9$	$5.74 \cdot 10^{14}$
6	$3.71 \cdot 10^7$	$2.37 \cdot 10^9$	$5.79 \cdot 10^{11}$	$6.03 \cdot 10^{17}$
7	$8.12 \cdot 10^8$	$1.04 \cdot 10^{11}$	$6.33 \cdot 10^{13}$	$6.93 \cdot 10^{20}$
8	$1.89 \cdot 10^{10}$	$4.84 \cdot 10^{12}$	$6.75 \cdot 10^{15}$	$8.66 \cdot 10^{23}$
9	$4.65 \cdot 10^{11}$	$2.39 \cdot 10^{14}$	$7.43 \cdot 10^{17}$	$1.17 \cdot 10^{27}$
10	$1.20 \cdot 10^{13}$	$1.35 \cdot 10^{16}$	$8.54 \cdot 10^{19}$	$1.70 \cdot 10^{30}$
11	$3.25 \cdot 10^{14}$	$8.24 \cdot 10^{17}$	$1.02 \cdot 10^{22}$	$2.63 \cdot 10^{33}$

Table 2. Simple random search. Minimum number of algorithm steps N for reaching an ε-vicinity of the D-optimal design with a level of certainty $1 - \gamma = 0.95$ for $X = [-1, 1]$, $m = n$.

n	$\varepsilon = 0.05$	$\varepsilon = 0.025$	$\varepsilon = 0.01$	$\varepsilon = 0.001$
2	4.04	4.01	4.00	4.00
3	9.02	8.96	9.00	8.99
4	18.44	18.38	18.45	18.45
5	36.12	36.21	36.24	36.56
6	69.11	68.90	69.01	67.07
7	130.92	130.00	129.71	114.17
8	241.38	242.00	220.59	184.26
9	442.86	445.90	357.21	283.98
10	810.81	888.16	561.84	423.94
11	1477.27	1831.11	864.41	610.21

Table 3. Efficiency of sampling from the generalized Δ^2-distribution vs simple random search in terms of the respective ratios of minimum number of algorithm steps, $1 - \gamma = 0.95$, $X = [-1, 1]$, $m = n$

distribution and taking the argument of the maximum value of $\Delta^2_{4,3}(Q)$, leads to the results showed in Table 4: we should choose two points at the interval boundaries and two points in the middle. When $n = 5 \equiv 2(\mathrm{mod}\, m)$, though, numerical approximation yields a non-unique solution (Table 5): the fifth "extra" point might be allocated at either end of the interval.

The simulation algorithm for the generalized Δ^2-distribution is designed to handle efficiently every set of orthonormalized functions $\varphi_1, \ldots, \varphi_m$. For instance, if we consider a non-traditional system by orthonormalizing 1, x, and e^x in $[0, 1]$, we can still get plausible results (Table 6).

N	x_1	x_2	x_3	x_4	Max
500	0.006376	0.471703	0.540401	0.991899	219/4
1000	0.004603	0.490417	0.521373	0.997730	217/4
2000	0.004025	0.493003	0.518318	0.999384	243/4
5000	0.003306	0.496601	0.508966	0.999937	243/4
10000	0.000772	0.502950	0.505861	0.999992	243/4

Table 4. Maximization of $f_{4,3}(Q)$ in $X = [0, 1]$: quadratic regression.

N	x_1	x_2	x_3	x_4	x_5	Max
500	0.028800	0.058415	0.482240	0.528539	0.973710	333/10
1000	0.009243	0.509697	0.527930	0.971588	0.987821	406/10
2000	0.008737	0.479380	0.482553	0.970622	0.992075	407/10
5000	0.004319	0.025813	0.494049	0.501784	0.998808	418/10
10000	0.002403	0.492002	0.512687	0.984694	0.989464	418/10

Table 5. Maximization of $f_{5,3}(Q)$ in $X = [0, 1]$: quadratic regression.

N	x_1	x_2	x_3	x_4	Max
500	0.006889	0.415713	0.448710	0.983057	0.06
1000	0.003014	0.524322	0.534542	0.991139	0.07
2000	0.003322	0.424577	0.505026	0.999384	0.08
5000	0.008751	0.471999	0.515507	0.998711	0.08
10000	0.000231	0.491175	0.498388	0.988331	0.09

Table 6. Maximization of $f_{4,3}(Q)$ in $X = [0, 1]$: $\varphi_1(x) = 1$, $\varphi_2(x) = \sqrt{3}(2x - 1)$, $\varphi_3(x) = \sqrt{\frac{2}{53e^2+292e-25}}\,(7e^x - 6(e + 1)x - 10e + 4)$.

Let us now consider a special two-dimensional case. Suppose $X = [0,1]^2$, $n = 7$, $m = 6$, and $\varphi_1, \ldots, \varphi_6$ are polynomials of not higher than quadratic order, i.e.

$$\varphi_1(x,y) = 1, \qquad \varphi_2(x,y) = \sqrt{3}\,(2x-1), \qquad \varphi_3(x,y) = \sqrt{3}\,(2y-1),$$

$$\varphi_4(x,y) = \sqrt{5}\,(6x^2 - 6x + 1), \qquad \varphi_5(x,y) = \sqrt{5}\,(6y^2 - 6y + 1),$$

$$\varphi_6(x,y) = 12xy - 6x - 6y + 3.$$

A preliminary notion about the allocation of the optimal design points might be extracted from Podkorytov (1975). In [10] he proved that for $m = n$, a convex region X, and a quadratic set of polynomials, the maximum of (2.1) is reached by selecting not more than one internal point for X, i.e. all the other points lie on the its border. Figure 1 illustrates this theoretical result numerically by sampling from the Δ^2- (Ermakov–Zolotukhin) distribution for $X = [-1,1]^2$, $m = n = 6$, $\varphi_1, \ldots, \varphi_6$ – orthonormalized in $L^2(X)$ Legendre polynomials of not higher than quadratic degree. For 500 simulations the approximation of the exact D-optimal design (the triangles in Figure 1) yields 6 points that lie on the boundaries of the optimization region $X = [-1,1]^2$. When we estimate the mode of the generalized Δ^2-distribution based on sample of size 1000, we detect a similar effect, the only "suspicious", i.e. possibly internal, point being $(-0.01, -0.71)$. We performed the same procedure by simulating 5000, 10000, 100000, and 200000 Δ^2-distributed vectors. As the results did not differ qualitatively, we present in Figure 1 the allocation of the D-optimal design just in the case $N = 5000$. Now the solution contains a distinct internal point $(-0.11, 0.17)$. The other 5 points lie on the boundary of $X = [-1,1]^2$ and seem to be concentrated at the vertices of the square. There are no theoretical results supporting the latter observation, but for sure Podkorytov's finding (not more than one internal point) is empirically supported.

The case $m = n$ pertains to the simulation (see [3]) of the (standard) continuous Δ^2-distribution. When $n > m$ we address its generalized version, and we can study a close special case: $n = 7$, $m = 6$. It turns out that a similar allocation of the optimum points is detected after simulating 2000 or more Δ^2-distributed vectors and taking the maximum argument. Indeed, Figure 2 shows that for $N \geq 2000$, we have not more than one internal point of the approximate D-optimal design.

In this case, however, it is quite clear that the problem has a non-unique solution. This phenomenon occurs when the objective function takes several values that are close to its global maximum. As a result, by sampling from the generalized Δ^2-distribution, we get a set of D-optimal designs corresponding to the maximum of $\Delta^2_{n,m}(Q)$ for the different sample sizes. It is important to classify the ties in this set. In this way, when we find one D-optimal design, we can transform it to the one in the set, which is most suitable for the specific research problem we work on.

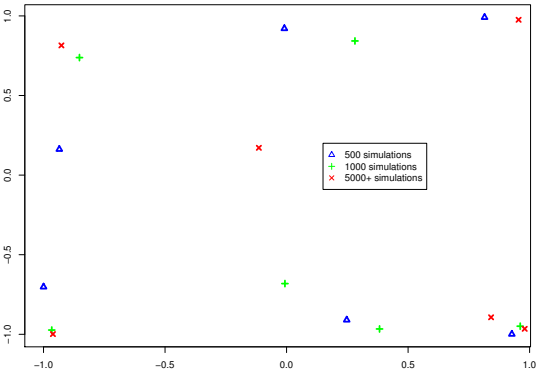

Figure 1. Numerical D-optimization of $f_{6,6}(Q)$ in $X = [-1, 1]^2$: quadratic regression.

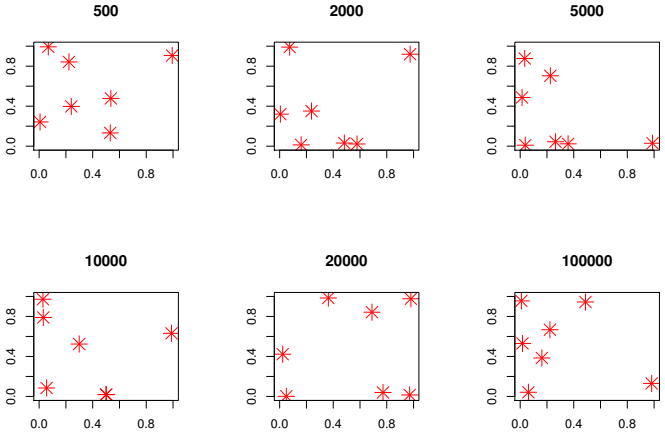

Figure 2. Maximization of $f_{7,6}(Q)$ in $X = [0, 1]^2$: quadratic regression. Number of generalized Δ^2 simulations: 500, 1000, 2000, 5000, 10000, 20000, 100000.

5. Discussion

In the problem of D-optimal design approximation simulating generalized Δ^2-vectors and assessing the mode of the resulting sample is justified. The structure of the generalized Δ^2-distribution is such that the determinant in (2.2) is equal to zero on a surface of dimension $s-1$, where $s = \dim X$. The latter is a property of the Vandermonde determinant and its multi-dimensional generalization. As a result, as s increases, the maximum in (2.2) will be higher.

In terms of sampling from the generalized Δ^2-distribution, the efficiency of simulation decreases significantly when n is substantially greater than m. If this is the case, however, continuous designs serve as very good estimates of the exact ones, and we can take advantage of the Kiefer–Wolfowitz theory.

6. Conclusion

The approach to global optimization suggested in this paper does not interfere with standard grid methods in the class of problems they are usually applied to. However, we have deliberately shown maximum estimation results pertaining to sampling from the generalized Δ^2-distribution. Focusing specifically on the problem of D-optimal design construction, we would like to point out that the non-uniqueness of its solution requires special attention. In accordance with that, we shall keep in the computation procedure not one, but rather a set of argument values, corresponding to the record values of $\Delta^2_{n,m}$. Difficulties related to the solution's non-uniqueness come out for almost all $n > m$ and especially when the region of interest X has high dimension. That is how we could interpret the substantial differences in the configuration of the D-optimal design points in Figure 2.

Certainly, the method of numerical maximum estimation shown in this article may be used in combination with other techniques like, for instance, differential evolution (see, for example, [11]). This is especially important while solving problems with multiple optima. Differential evolution provides a number of analytical and purely visual means for efficient searching strategies.

References

1. K. Bogues, C.R. Morrow, and T.N.L. Patterson. *An Implementation of the Method of Ermakov and Zolotukhin for Multidimensional Integration and Interpolation*, Numer. Math. 37 (1981), pp. 49–60.
2. S.M. Ermakov. *Die Monte-Carlo Methode und verwandte Fragen*, Deutsch. Verlag Wissenschaft, 1975.
3. S.M. Ermakov and T.I. Missov *Simulation of the* Δ^2*-distribution*, Vestnik SPbGU 4 (2005), pp. 123–140.
4. S.M. Ermakov and A.A. Zhigljavsky *Mathematical Theory of Optimal Design*, Nauka, 1987.

5. S.M. Ermakov and V.G. Zolotukhin. *Polynomial approximations and the Monte Carlo Method* , Theor. Probability Appl. 5 (1960), pp. 428–431.

6. V.V. Fedorov *Theory of Optimal Experiments* translated and edited. In: W.J. Studden and E.M. Klimko (editors), Academic Press, New York, 1972.

7. J. Kiefer and J. Wolfowitz. *Optimum Designs in Regression Problems*, The Annals of Mathematical Statistics 30 (1959), pp. 271–294.

8. T.I. Missov. *Integral Evaluation Using the Δ^2-distribution. Simulation and Illustration*. Monte Carlo Methods and Applications 13 (2007), pp. 219–225.

9. T.I. Missov and S.M. Ermakov. *Simulation of the Δ^2-distribution. The discrete case*. Proceedings of the 5th St. Petersburg Workshop on Simulation, pp. 499–502, 2005.

10. A.N. Podkorytov. *On the Properties of D-optimal Designs for Quadratic Regression*, Vestnik LGU 2 (1975), pp. 163–166.

11. K.V. Price, J. Lampinen, and R. Storn. *Differential Evolution*, Springer-Verlag, Berlin, 2005.

12. E. Riccomagno, R. Schwabe, and H.P. Wynn *Lattice-Based D-Optimum Design for Fourier Regression*, The Annals of Statistics 25 (1997), pp. 2313–2327.

13. J. Schur. *Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten*, Math. Zeit. 1 (1918), pp. 377–402.

14. H.P. Wynn. *The Sequential Generation of D-Optimum Experimental Designs*, The Annals of Mathematical Statistics 41 (1970), pp. 1655–1664.

15. A.A. Zhigljavsky and A.G. Zilinskas. *Stochastic Global Optimization*, Ser. Springer Optimization and Its Applications, Vol. 9, Springer-Verlag, Berlin, 2008.

Received July 14, 2008; revised December 15, 2008

Author information

Trifon I. Missov, Department of Stochastic Simulation, Saint Petersburg State University, and Max Planck Institute for Demographic Research, Germany.
Email: Missov@demogr.mpg.de

Sergey M. Ermakov, Head of the Department of Stochastic Simulation, Saint Petersburg State University, Russia.
Email: Sergej.Ermakov@gmail.com