Data Squashing for Tail Inference in the Generalized Pareto Distribution

Xeni K. Dimakos*

Abstract

Data squashing was proposed by DuMouchel et al. (1999) as a tool for reducing a massive data set to a considerably smaller and more manageable one, that represents the original data set well for the purpose of inference or prediction. Data squashing is not a sub-sampling technique and the new records are pseudo data points generated on the space spanned by the original data set. The idea is to find these new points so that a certain set of empirical moments of the squashed data set is equal to the same set of empirical moments of the original data set.

This paper reviews data squashing and applies the method to data from the generalized Pareto distribution (GPD). This family of distributions is of particular interest because it includes heavy tailed distributions. As a traditional data reduction technique such as simple random sampling can be inadequate in representing heavy tails and inference based on such samples can suffer from poor accuracy, it is of interest to see if data squashing can improve upon sub-sampling for heavy tailed data.

As a first step in exploring the properties of data squashing for heavy tailed data, we consider one GPD which has finite third order moment and one GPD which has finite expectation but infinite higher order moments. In a simulation experiment we explore how this affects data squashing compared with sub-sampling with respect to maximum likelihood estimation of the tail parameter in the GPD.

*Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway. Email: xeni@math.uio.no

Our results indicate that data squashing works well and show that the variability of the ML estimates are considerably smaller for data squashing than for the applied sub-sampling technique. We also see improvements in the bias, although the difference here is smaller.

Key words: Data squashing, generalized Pareto distribution, heavy tails, massive datasets, moment matching, Taylor expansion

1 Introduction

During the last decade the general progress of information technology has led to the development of powerful data base and data warehouse products that are now widely used. This development encourages to collect data for the purpose of extracting information that can assist in decision making, strategic planning or system monitoring, which is often referred to as data mining (DM) or knowledge discovery in data bases (KDD). The data may be historic or collected and processed in real time. The data bases are growing in size, and often traditional techniques for analysis and visualization are either not applicable or become severely CPU demanding. The standard approach to analysis of these massive data set have been to scale up the computer hardware and software in terms of increasing memory and processing capacity and by designing or choosing statistical methods that are less sensitive to the amount of data.

Data squashing represents an alternative way to approach massive data sets. Rather than increasing the computing power and choosing statistical methods that can handle large amounts of data, the data set of interest is represented by a smaller data set that can be used in combination with any advanced statistical method and software. From a modeling perspective this gives considerable flexibility. Not only one, but several models can be fitted and it is feasible to use methods for model checking and diagnostics. When possible, it is appealing to use a squashed data set for exploratory analysis and for choosing an appropriate model and then apply this final model to the full data set.

Naturally, conventional sampling techniques such as simple random sampling or stratified random sampling, provide the same reduction in data size as data squashing and are computationally faster and simpler. However, as demonstrated in DuMouchel et al. (1999) a squashed data set can be superior to a data set generated by simple

random sampling in terms of improved accuracy in inference, which compensates for the increase in computing time and effort.

In this paper we review data squashing as proposed in DuMouchel et al. (1999). Our aim is to study data squashing for heavy tailed data through simulation experiments. In a data reduction context, heavy tailed data are interesting because sub-samples often need to be large in order to represent the full data set well. With a massive and heavy tailed data set, one may experience a trade-off between having a sub-sample of manageable size and having a sub-sample that yields the desired accuracy when used for inference. We have chosen to use the generalized Pareto distribution (GPD) for our studies and we focus on its tail parameter. The maximum likelihood estimate of this tail parameter has no explicit form and is found by numerical optimization.

We apply data squashing to two data sets sampled from the generalized Pareto distribution and compare data squashing with stratified random sampling for maximum likelihood estimation of the tail parameter. The first of the test distributions has finite moments up to order three, while the second has finite expectation, but infinite moments of higher order. It is of interest to see if data squashing is sensitive to these features.

As data squashing is very time consuming we have not been able to make the simulation study as extensive as we have wished, and further work is required to fully understand and verify the properties of data squashing for heavy tailed data. Furthermore, our results are based on an implementation of data squashing which has yet to be tested and assessed completely. However, the results we have obtained indicate that data squashing improves upon stratified random sampling. Most importantly, the variability in the ML estimates is considerably smaller for data squashing than for stratified random sampling. Secondly, there is a tendency toward improvement in the bias of the ML estimates based on the squashed data sets, but the difference is smaller than with respect to the variability. The best results are obtained for the test distribution with the least heavy tail, indicating that an increase in tail heaviness does have some influence on the properties of data squashing.

2 Data Squashing

In our presentation of data squashing we consider a massive data set with records $\mathbf{X}_i = (X_{i1}, \ldots, X_{iQ}), i = 1, \ldots, N$, where Q is the fixed number of variables in each

record. In our application of data squashing we will only consider scalar data, i.e. $Q = 1$. The data set has a simple flat structure and can be represented as a $N \times Q$ matrix. The data set is massive in the sense that N is very large, typically of the order 10^5 or 10^6, while Q is moderate. Of course, if the statistical methods we are interested in applying are computationally demanding, then also with a smaller data set one can encounter computational difficulties. For clarity of presentation we will assume that the data set contains continuous variables only. The modifications to include categorical variables are minor and described in DuMouchel et al. (1999).

The aim of data squashing is to generate a new and managable data set of records $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jQ})$, $j = 1, \ldots, M$ with $M \ll N$. Non-negative weights $w_1, \ldots, w_M$ are associated with each record and generated so that $\sum_{j=1}^{M} w_j = N$. In contrast to a data set generated by simple random sampling or other sampling techniques, the squashed data set is not a sub set of the massive data set. As will become apparent, the squashed points are new pseudo points generated on the space spanned by the full data set. The idea is that the squashed data set is used for inference in place of the original data set, for instance in doing maximum likelihood estimation or fitting regression models. However, it is important that the applied methods incorporate the weights associated with the squashed points.

Generating the squashed points and weights is treated as a problem of determining $M(Q+1)$ unknown variables. The aim is that the empirical moments of the squashed data set approximate the empirical moments of the massive data set. More precisely, we want $K \geq M(Q+1)$ empirical weighted moments of the squashed data set to be equal to or match the corresponding empirical unweighted moments of the full data set, that is

$$\sum_{j=1}^{M} w_j \prod_{q=1}^{Q} (Y_{jq} - a_q)^{p_{kq}} = \sum_{i=1}^{N} \prod_{q=1}^{Q} (X_{iq} - a_q)^{p_{kq}} \quad k = 1, \ldots, K. \tag{1}$$

We are allowed to refer to the equations (1) as a matching of moments because we will use $\sum_{j=1}^{M} w_j = N$ which means that the scaling factor of empirical moments cancels. Furthermore, we define an empirical moment about $\mathbf{a} = (a_1, \ldots, a_Q)$ through exponent vectors $\mathbf{p}_1, \ldots, \mathbf{p}_K$ of non-negative integers. In principle, such an exponent vector can be any arbitrary sequence of non-negative integers. A natural choice is to start with vectors of zeros and ones and let the integers increase with k. One possibility is to take $\mathbf{p}_1 = (0, \ldots, 0), \mathbf{p}_2 = (1, 0, \ldots, 0), \ldots, \mathbf{p}_{Q+1} = (0, \ldots, 0, 1), \mathbf{p}_{Q+2} = (2, 0, \ldots, 0) \ldots,$

a.s.o.

As a simple illustration consider (1) for scalar data $X_1, \ldots, X_N$. With $p_k = k - 1$ and with $a = 0$ the equations in (1) reduce to

$$\sum_{j=1}^{M} w_j Y_j^{k-1} = \sum_{i=1}^{N} X_i^{k-1} \quad k = 1, \ldots, K. \tag{2}$$

If we want to generate only one squashed point and weight, we need $K = 2$ equations. This yields $Y = \overline{X}$ and $w = N$.

DuMouchel et al. (1999) suggest to determine the pseudo points $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ and weights $w_1, \ldots, w_M$ by minimizing

$$S(\mathbf{Y}, \mathbf{w}) = \sum_{k=1}^{K} u_k \left(\sum_{i=1}^{N} \prod_{q=1}^{Q} (X_{iq} - a_q)^{p_{kq}} - \sum_{j=1}^{M} w_j \prod_{q=1}^{Q} (Y_{jq} - a_q)^{p_{kq}} \right)^2, \tag{3}$$

for optimization weights u_k, $k = 1, \ldots, K$ under the constraint that the weights should be positive and the points should not extrapolate, i.e. $w_j \geq 0$, $\forall j$ and $\min_i X_{iq} \leq Y_{jq} \leq \max_i X_{iq}$, $j = 1, \ldots, M, q = 1, \ldots, Q$. Hence, the original problem of finding the roots in (1) is transformed to an optimization problem. The optimization weights $u_k, k = 1, \ldots, K$ determine which moments are approximated most accurately and the demand to match moments exactly is abandoned. Typically there is no unique global minimun and several local minima.

DuMouchel et al. (1999) propose a Newton-Raphson procedure that uses second order derivatives to minimize (3). A logistic transformation of the unknowns is used to maintain the constraints on the squashed points and weights. DuMouchel et al. (1999) suggest to take $u_k = 1000$ for the moments of order 0 and 1 defined by $\mathbf{p}_1 = \mathbf{0}, \mathbf{p}_k = (p_{k(k-1)} = 1, p_{kr} = 0, r \neq k-1), k = 2, \ldots, Q+1$ and for the pure squares defined by $\mathbf{p}_k = (p_{k(k-Q-1)} = 2, p_{kr} = 0, r \neq k-Q-1), k = Q+2, \ldots, 2Q+1$. For the second order cross terms and for the higher order terms, the weights decrease with increasing order of the expansion and are normalized so that $\sum_{k=2Q+2}^{K} u_k = 1$. Observe that one does not perform a constrained optimization with $\sum_{j=1}^{M} w_j = N$ but includes this criterion as a term in the objective function. DuMouchel et al. (1999) analyse the properties of data squashing with respect to CPU demands and conclude that the running time increases linearly in Q and as a result of choosing $M = \log_2 N$ it increases in $(\log_2 N)^2$.

The moment matching criterion (1) may also be derived using a Taylor expansion argument. Assume that the original N data records are iid from a density $f(\mathbf{x}_i; \boldsymbol{\theta})$ with

log-likelihood $l_x(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(\mathbf{x}_i; \boldsymbol{\theta})$. Recall that the mth degree Taylor polynomial $P_m(\mathbf{x})$ for a function $f : \mathbb{R}^Q \to \mathbb{R}$ around a point $\mathbf{a} = (a_1, \ldots, a_Q)$ is given by

$$P_m(\mathbf{x}) = f(\mathbf{a}) + \sum_{r=1}^{m} \frac{1}{r!} \sum_{q_1=1}^{Q} \cdots \sum_{q_r=1}^{Q} \frac{\partial^r f(\mathbf{a})}{\partial x_1 \cdots \partial x_q} (x_{q_1} - a_{q_1}) \cdots (x_{q_r} - a_{q_r}).$$

It follows that the Taylor expansion of the log-likelihood in the vicinity of a point $\mathbf{a} = (a_1, \ldots, a_Q)$ is

$$l_x(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) \approx \sum_{k=1}^{K} g_k \prod_{q=1}^{Q} (x_q - a_q)^{p_{kq}}. \tag{4}$$

Here g_k depends on $\boldsymbol{\theta}$ and on $\log f$ through its derivatives calculated in $\mathbf{a}$ but not on the data. If the Taylor expansion is taken up to order m then $K = \sum_{r=0}^{m} \binom{Q+r-1}{r}$ and $\mathbf{p}_k, k = 1, \ldots, K$ are all possible vectors of non-negative integers such that $\sum_{q=1}^{Q} p_{kq} \leq m$. In (4) the first term has $\mathbf{p}_1 = (0, 0, \ldots, 0)$ and corresponds to the constant term $f(\mathbf{a})$ in the expansion. The sum of the next Q terms in (4) corresponds to the Taylor expansion of order one with $\mathbf{p}_k = (p_{k(k-1)} = 1, p_{kr} = 0, r \neq k-1), k = 2, \ldots, Q+1$. More generally, the rth order of the Taylor approximation is represented by $\binom{Q+r-1}{r}$ terms in (4). We want the squashed data set to represent as much as possible of the information on $\boldsymbol{\theta}$ contained in the full data. Hence, we want the weighted log-likelihood of the squashed data set $l_y(\mathbf{y}_1, \ldots, \mathbf{y}_M; \boldsymbol{\theta}) = \sum_{j=1}^{M} w_j \log f(\mathbf{y}_j; \boldsymbol{\theta})$ to approximate the full likelihood. By equating the Taylor expansions of the log-likelihood of the full and squashed data sets and by changing the order of summation we obtain

$$\sum_{k=1}^{K} g_k \sum_{j=1}^{M} \prod_{q=1}^{Q} w_j (Y_{jq} - a_q)^{p_{kq}} = \sum_{k=1}^{K} g_k \sum_{i=1}^{N} \prod_{q=1}^{Q} (X_{iq} - a_q)^{p_{kq}}.$$

Since this is to hold for any model f and hence for any g_k, the equality needs to hold term by term. This is exactly the moment matching criterion (1) with $\mathbf{p}_k, k = 1, \ldots, K$ determined by the Taylor expansion.

An important feature of data squashing is that the massive data set can be grouped into regions and data squashing performed independently for each region. DuMouchel et al. (1999) apply two different techniques for grouping the data in regions: hyper-rectangles which suffer from a curse of dimensionality and are suitable only when Q is small and data spheres (Johnson and Dasu, 1998) which can be used with high

ξ	$\mathrm{E}(X)$	$\mathrm{E}(X^2)$	$\mathrm{E}(X^3)$	$\mathrm{E}(X^4)$
0.25	4/3	16/3	64	∞
0.5	2	∞	∞	∞

Table 1: *Moments of the GPD with* $\xi = 0.25$ *and* $\xi = 0.5$ *for* $\beta = 1$.

dimensional data. An advantage of grouping the data is that the Taylor approximation is improved for each region, compared with an approximation on the full data set. Also, grouping allows to make use of parallel computing in the minimization.

We may summarize data squashing in three steps: (i) An optional grouping of the data and determination of how many squashed points that should be generated for each region. (ii) Calculation of empirical moments of the massive data set. (iii) The optimization that finds the squashed points and weights for each region.

Data squashing is a novel techique and to our knowledge the literature is limited to Madigan et al. (2000), Owen (1999) and Berg et al. (2000). Madigan et al. (2000) present a version of data squashing that uses the density function of the original data records, so that the squashed data set may only be used for inference in the assumed model. Owen (1999) suggests to combine the ideas of data squashing and sub-sampling by determining the weights for the records in a sub-sample by matching moments, while Berg et al. (2000) give an overview of data squashing and present a collection of ideas for further research.

3 The Generalized Pareto Distribution

In extreme value modeling the generalized Pareto distribution (GPD) is used to model excesses over high thresholds. Embrechts et al. (1997, Sections 3.4 and 6.5) cover the field and give a thorough treatment of the GPD and its properties. A random variable X that is distributed according to a GPD with parameters $\xi \neq 0$ and $\beta > 0$ has density function

$$f(x) = \frac{1}{\beta}(1 + \xi\frac{x}{\beta})^{-1/\xi - 1},$$

with support $x \geq 0$ if $\xi \geq 0$ and $x \in [0, -\beta/\xi]$ if $\xi < 0$. For $\xi = 0$ the GPD is the exponential distribution with expectation $1/\beta$. While β is a scale parameter, ξ determines the tail of the distribution. The larger ξ, the more heavy is the tail.

The expectation is finite if $\xi < 1$ and the rth moment is finite if $r < 1/\xi$. Table 1 summarizes moments about zero for the distributions that will be used in the simulation experiments in Section 4. Figure 3 shows the corresponding density functions.

When $\mathbf{X} = (X_1, \ldots, X_N)$ are iid from a GPD, the log-likelihood is

$$l(\xi, \beta; \mathbf{X}) = -n \log \beta - (\frac{1}{\xi} + 1) \sum_{i=1}^{N} \log(1 + \frac{\xi}{\beta} X_i). \tag{5}$$

In Section 4 we will do maximum likelihood estimation on a squashed data set $\mathbf{Y} = (Y_1, \ldots, Y_M)$ with weights $\mathbf{w} = (w_1, \ldots, w_M)$ assumed to follow a GPD. The weighted log-likelihood is

$$l(\xi, \beta; \mathbf{Y}, \mathbf{w}) = -\sum_{j=1}^{M} w_j \log \beta - (\frac{1}{\xi} + 1) \sum_{j=1}^{M} w_j \log(1 + \frac{\xi}{\beta} Y_j). \tag{6}$$

For unweighted and weighted data the maximum likelihood estimates of ξ and β are obtained by numerical maximization of (5) and (6) respectively. In the simulation experiments in Section 4 we focus on the tail parameter ξ and explore how the accuracy of the maximum likelihood estimate of ξ based on sub-sampled and squashed data sets varies with the heaviness of the tail.

4 Simulation Experiments

In our simulation experiment we do data squashing for two data sets of $N = 100\,000$ records sampled from the GPD with tail parameters $\xi = 0.25$ and $\xi = 0.5$ and fixed scale parameter $\beta = 1$. The true densities are shown in Figure 3. The size of these data sets by no means call for data squashing as any software easily handles this amount of data. However, we believe the findings to be of interest for larger and multidimensional data sets with similar characteristics, where reducing the data would be needed.

The intension of the study is to explore the properties of data squashing for our two test distributions that differ in tail heaviness, with respect to accuracy in the obtained ML estimate of the tail parameter. We compare data squashing with stratified random sampling. In DuMouchel et al. (1999) data squashing is compared with simple random sampling rather than stratified random sampling as done here. In our opinion data squashing and simple random sampling can not be compared on an equal basis as the grouping done in data squashing induces a spread in the generated points that is not achieved without stratification in sub-sampling.

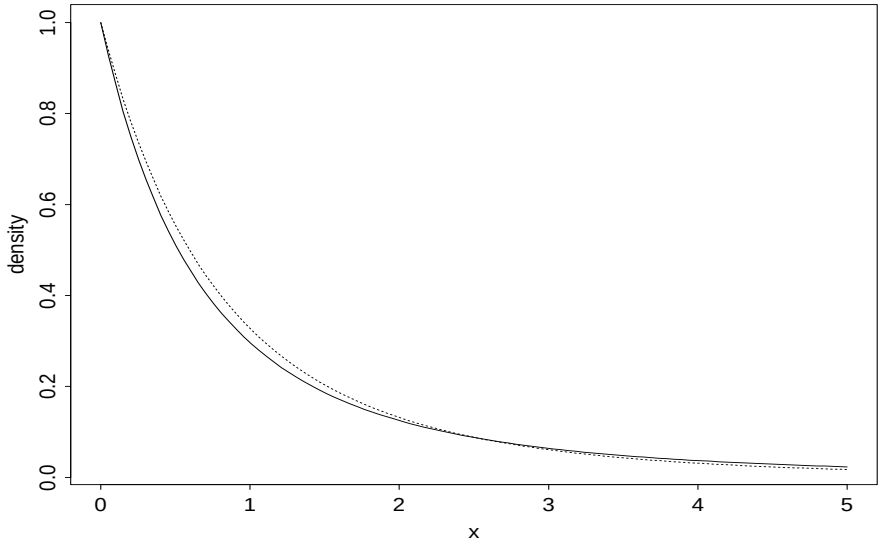


Figure 1: *The density function for the generalized Pareto distribution with* $\xi = 0.25$ *(dashed line) and* $\xi = 0.5$ *(solid line) for* $\beta = 1$.

Application of a global optimization routine, such as for instance simulated annealing, in the data squashing would ensure that the squashed data set would correspond to a global minimum in (3). Unfortunately, these algorithms require too much CPU and are not applicable in practice. In a realistic optimization setting one must settle for a computationally feasible local optimization algorithm. Hence, the squashed data set will depend on the type of optimization algorithm used and its initialization.

We applied a C-language implementation of data squashing that is currently being developed at the Norwegian Computing Center. This implementation uses a Newton-Raphson algorithm with multiple starting points (we used 1000) to find the squashed data set. As default values for the optimization weights in (3) for the constraint on the weights and the first and second order moments this implementation takes $u_k = 1$ for $k = 1, 2, 3$ respectively. For the moments of higher order the default values are $u_k = 0.5^{k-3}/c$ where c is chosen so that $\sum_{k=4}^{K} u_k = 0.001$. This is basically the same strategy as applied in DuMouchel et al. (1999), but with the objective function (3) scaled by 0.001. However, DuMouchel et al. (1999) do not specify how they let the weights decrease for $k \leq 4$, nor do they give any recommendation on how this should be done. Furthermore, we used 500 as an upper limit of the number of iterations in

the optimization.

To simulate data from the GPD and to do maximum likelihood estimation in model (5) we used the S-Plus software of McNeil (2000). Some minor adjustments were required to use the weighted log-likelihood (6) for the squashed data.

To average out the effect of the random initialization of the Newton-Raphson procedure, we repeated the squashing $B = 50$ times and did maximum likelihood estimation for each of the obtained data sets. From this we obtained ML estimates $\widehat{\xi}_1, \ldots, \widehat{\xi}_B$ and found the mean point estimate $\widehat{\xi}_{\text{DS}} = \sum_{b=1}^{B} \widehat{\xi}_b / B$ and the estimated standard deviation $\widehat{\sigma}_{\text{DS}} = (\sum_{b=1}^{B} (\widehat{\xi}_b - \widehat{\xi}_{\text{DS}})^2 / (B-1))^{1/2}$. The estimated mean squared error $\widehat{\gamma}_{\text{DS}} = \sum_{b=1}^{B} (\widehat{\xi}_b - \widehat{\xi}_{\text{ORG}})^2 / B$ is found with respect to the ML estimate of the full data set $\widehat{\xi}_{\text{ORG}}$ and takes into account both the estimated bias and the variability of the ML estimates. For the two test data set $\widehat{\xi}_{\text{ORG}}$ was found to be 0.247 and 0.497 respectively, slightly below the true values.

Data squashing was applied to the each test data set for two different grouping strategies. With the first approach, the data was grouped into 10 bins of 10 000 records according to the percentiles $0.1, 0.2, \ldots, 0.9$. The second approach used 20 regions of 5 000 records each, determined by using the percentiles $0.05, 0.1, \ldots, 0.9, 0.95$ as bin borders. For each of these two grouping strategies we generated 1, 3 or 5 squashed points in each region. Hence, we applied a total of 6 different squashing techniques. The squashed data sets had $M = \{10, 30, 50\}$ and $M = \{20, 60, 100\}$ data points for the two grouping strategies respectively. Following (1) and (2) we needed to match 1, 5 and 9 empirical moments as well as the constraint on the weights to generate 1, 3 and 5 points and weights per region. Hence, some of the empirical moments that are matched correspond to theoretical moments that are infinite, except for when only one point is found per region.

The program was run on a Sun Ultra 30 with an UltraSPARC-II 300 MHz processor. When only one point is generated per region as with $M = 10$ and $M = 20$, the region mean is always a unique global minimum. For each region the squashing took seconds and there is no need to repeat the squashing as exactly the same data set is found each time. The minimization to find 3 and 5 points in each region is more difficult as (3) is then a 6 and 10 dimensional surface. To generate 3 and 5 points per region using multiple starting points took approximately 4.75 and 9 minutes per region.

To each of the applied squashing techniques the corresponding stratified random sampling technique used the same data regions and the same number of points were

$\xi = 0.25$

	10	30	50	20	60	100
MEAN	0.223	0.243	0.254	0.246	0.246	0.249
SD	–	0.008	0.011	–	0.003	0.003
MSE	5.91e-04	8.1e-05	1.63e-04	2e-06	1.1e-05	1.2e-05

$\xi = 0.5$

	10	30	50	20	60	100
MEAN	0.557	0.523	0.475	0.540	0.513	0.497
SD	–	0.016	0.016	–	0.013	0.018
MSE	0.0036	9.39e-04	7.24e-04	0.0019	4.29e-04	3.18e-04

Table 2: *Maximum likelihood estimation of ξ based on the 6 squashing techniques for the full data sets with $\xi = 0.25$ and $\xi = 0.5$. For each squashing technique and for each value of ξ a total of 50 squashed data sets were collected to assess the performance of the procedure. The tables show the estimated mean, standard deviation and mean squared error of the ML estimates for the resulting squashed data sets. The left and right hand side of the tables separate the grouping strategy with 10 and 20 regions respectively.*

sampled with replacement from each region. To quantify the bias and variability associated with stratified random sampling $B = 10000$ subsamples were generated for each sub-sampling approach and test data set. The estimated mean $\widehat{\xi}_{\text{SRS}}$, standard deviation $\widehat{\sigma}_{\text{SRS}}$ and mean squared error $\widehat{\gamma}_{\text{SRS}}$ were found as for data squashing.

The estimated mean, standard deviation and mean squared error of the ML estimates for the data sets generated by data squashing and stratified random sampling are shown in Tables 2 and 3 respectively. Figure 2 shows the true log-likelihood for $\xi = 0.25$ and 10 log-likelihoods for data sets generated by data squashing and sub-sampling. Figure 3 shows log-likelihoods for $\xi = 0.5$. Observe that to ease the comparison of the log-likelihoods in Figures 2 and 3 the log-likelihoods were shifted upwards and plotted within the same horizontal coordinates. Also, the log-likelihood for the stratified random samples were found by assigning equal weights of N/M to each sampled data point.

$\xi = 0.25$

	10	30	50	20	60	100
Mean	0.115	0.191	0.210	0.201	0.227	0.233
Sd	0.267	0.174	0.134	0.143	0.093	0.074
Mse	0.089	0.033	0.019	0.023	0.009	0.006

$\xi = 0.5$

	10	30	50	20	60	100
Mean	0.394	0.455	0.469	0.463	0.483	0.487
Sd	0.287	0.187	0.145	0.156	0.099	0.078
Mse	0.093	0.037	0.022	0.025	0.010	0.006

Table 3: *Maximum likelihood estimation of ξ based on stratified random sampling. For each full data set with $\xi = 0.25$ and $\xi = 0.5$ a total of 10 000 independent stratified samples were generated for each of the 6 combinations of grouping and data reduction, and the ML estimate of ξ found for each sample. The tables show the estimated mean, standard deviation and mean squared error based on these 10 000 estimates. The left and right hand side of the tables separate the grouping strategy with 10 and 20 regions respectively.*

The standard deviation of the ML estimates obtained using asymptotic normal theory are found to be approximately 0.004 and 0.005 for $\xi = 0.25$ and $\xi = 0.5$. These results are similar for data squashing and stratified random sampling.

Comparing Tables 2 and 3 we see that the results for data squashing and stratified random sampling differ with respect to both the bias and the variability of the ML estimates. The most marked difference between the two techniques is found for the variability of the ML estimates, which is considerably smaller for data squashing for both test data sets and for all the applied squashing techniques. For $\xi = 0.25$ the estimated variability associated with the squashing procedure is of the same order as the asymptotic variability associated with the maximum likelihood estimation. Figure 2 shows that the log-likelihoods of the squashed data sets approximate the true log-likelihood much better than the log-likelihoods of the sub-sampled data sets for $\xi =$

0.25. For $\xi = 0.5$ (Figure 3) the pattern is similar, although the log-likelihoods are slightly more spread than for $\xi = 0.25$.

The results also show that there is a tendency for the bias to be smaller when data squashing is applied, although the improvement is not of the same order as for the variability. For $\xi = 0.25$ the bias is consistently lower for data squashing, while for $\xi = 0.5$ the results for data squashing and stratified random sampling are quite close, and for $M = 20$ and $M = 60$ the point estimate based on stratified random sampling has a smaller bias. For both techniques, $\widehat{\xi}_{\text{ORG}}$ is covered by the interval $(\widehat{\xi} - 2\widehat{\sigma}, \widehat{\xi} + 2\widehat{\sigma})$ or an even shorter interval.

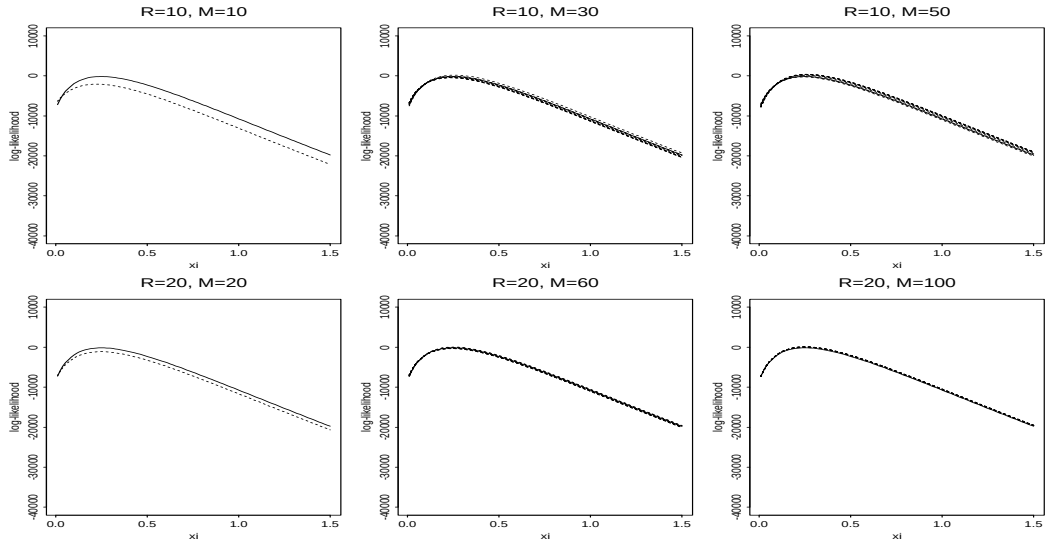
It should be noted that for stratified random sampling some of the ML estimates were negative for $\xi = 0.25$ and $M = 10, 20, 30, 50$ as well as for $\xi = 0.5$ and $M = 10, 30$. For $\xi = 0.25$ and $M = 10$ this occurred for approximately every third sampled data set, which explains the resulting low value in Table 3 for this strategy.

For reference it should be noted that simple random sampling does much worse than stratified random sampling. The bias of the estimates is larger and correspond to an effect in the second decimal of the ML estimates, while the estimated standard deviation is considerably larger, in fact it is approximately doubled for $M = 60$ and $M = 100$.

We compared the variability associated with data squashing in the lower region (the 5 000 and 10 000 smallest values with 20 and 10 regions respectively) and in the the tail region (the 5 000 and 10 000 largest values) with the following experiment. For the 50 collected squashed data sets the squashed points in the lower region and tail region were sorted in increasing order. The ratio of the mean and standard deviation of the first, second, third, etc. points and were found, and the procedure repeated for the associated weights. The results indicated that the points and weights in the tail region were more variable than in the lower region. The variability in the tail region was larger for $\xi = 0.5$ than for $\xi = 0.25$, while the variability of the points and weights in the lower region was unaffected by an increase in ξ. This indicates that there could be a variability associated with data squashing that increase with the tail heaviness. Also the results of Table 2 shows that data squashing works best for the test distribution with the least heavy tail.

As data squashing is very time consuming, we have been unable to make the simulation study as extensive as we would have liked and we feel that further work is needed to fully explore the properties of data squashing for heavy tailed data. Our

DATA SQUASHING: $\xi = 0.25$



STRATIFIED RANDOM SAMPLING: $\xi = 0.25$

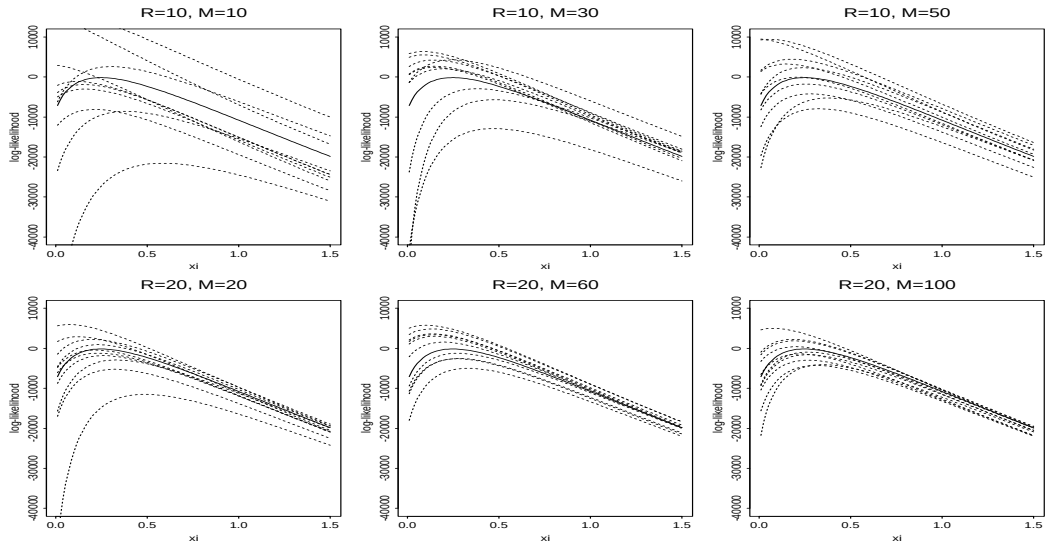
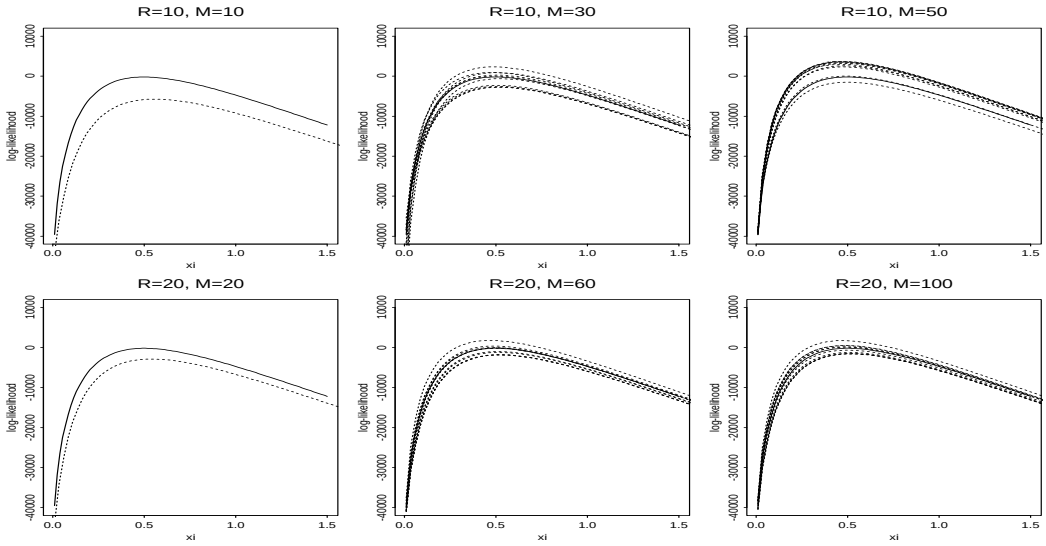


Figure 2: *The marginal log-likelihoods for* $\xi = 0.25$ *for 10 squashed data sets (upper panel) and 10 independent stratified random samples (lower panel) plotted in* $\widehat{\beta}$ *found from each reduced data set. In each plot the solid line shows the log-likelihood of the full data set. The number of regions* (R) *and size of the reduced data set* (M) *is indicated in the title of each plot.*

DATA SQUASHING: $\xi = 0.5$



STRATIFIED RANDOM SAMPLING: $\xi = 0.5$

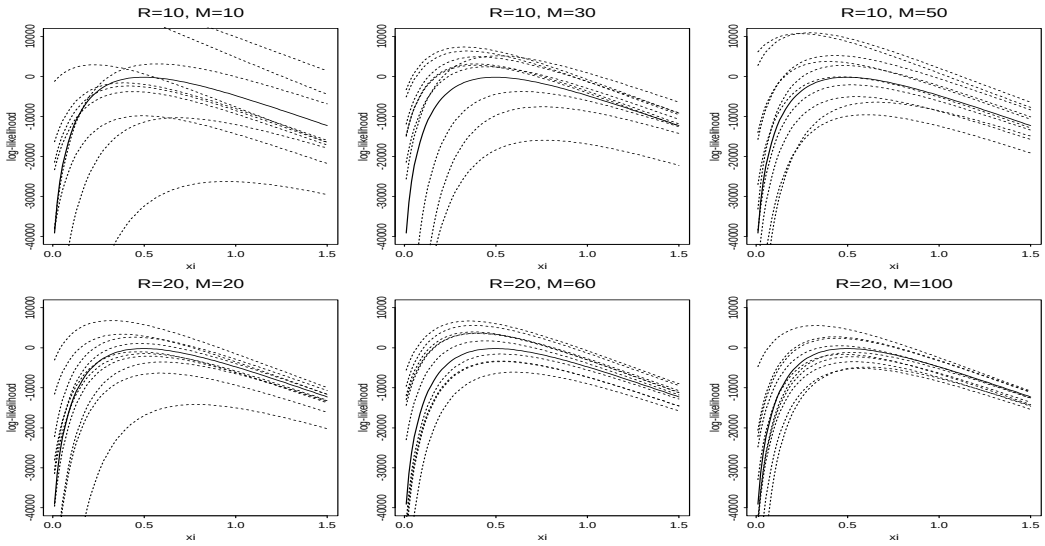


Figure 3: *The marginal log-likelihoods for $\xi = 0.5$ for 10 squashed data sets (upper panel) and 10 independent stratified random samples (lower panel) plotted in $\widehat{\beta}$ found from each reduced data set. In each plot the solid line shows the log-likelihood of the full data set. The number of regions (R) and size of the reduced data set (M) is indicated in the title of each plot.*

simulation experiment is limited to only one full data set of 100 000 records for each of the two tail parameters $\xi = 0.25$ and $\xi = 0.5$, and we compared data squashing with stratified random sampling for this data set only. To average out the variation in the ML estimates induced by the sampling of the data, more than one full data set should have been created and analysed. We performed a small test which shows that the results would be similar for other data sets. We generated 15 independent data sets of 100 000 records for $\xi = 0.25$. The first column of Table 4 shows the ML estimates for the full data sets and confirm that they do not vary significantly. For each of these data sets a total of 5 (as opposed to 50 in our full experiment) squashed set were found for the squashing technique that applied 10 regions and 3 points per region. The ML estimate for each of the 5 squashed sets were found and their average is listed in the second column of Table 4. Observe that the estimates in Tables 2 and 4 are not directly comparable as the estimates in Table 4 are based on fewer squashed data sets. The third column of Table 4 shows the ML estimate based on 10 000 stratified random samples. The results presented in Table 4 show that the ML estimates obtained from data squashing and stratified random sampling do not vary much over the data sets. The results indicate that data squashing is perhaps slightly more sensitive to the data set than stratified random sampling. Both techniques consequently underestimate the tail parameter, except for two cases of data squashing. The bottom line of the table shows the mean value over the 15 data sets and indicates that the difference in the bias of the two techniques might be somewhat smaller than for our single data set.

We experienced occasional problems in realizing the constraint on the weights of the squashed data set. Observe that we did not do constrained optimization with $\sum_{j=1}^{M} w_j = N$. Instead, we followed DuMouchel et al. (1999) and included this criterion as a term in the objective function (3) and we used $\sum_{j=1}^{M} w_j - N$ to measure convergence. In particular, failure to reach $\sum_{j=1}^{M} w_j = N$ was an apparent feature when we tested data squashing for two other distributions with even heavier tails than our two test distributions. Furthermore, our preliminary findings suggest that this problem only occurs for squashing in the tail region. As realizing $\sum_{j=1}^{M} w_j = N$ is favoured in the optimization through the specific optimization weight u_1 we used, we interpret these problems as a sign of difficulties in finding a local minimum in the optimizer and as an indication of that we might have used a too strict limit for the maximum allowed number of iterations in the optimization. This suggests that our study needs to be verified and supplemented by experiments in which the optimization is monitored

$\widehat{\xi}_{\text{ORG}}$	$\widehat{\xi}_{\text{DS}}$	$\widehat{\xi}_{\text{SRS}}$
0.247	0.243	0.191
0.256	0.251	0.200
0.253	0.251	0.197
0.255	0.242	0.199
0.248	0.238	0.191
0.247	0.245	0.191
0.252	0.246	0.197
0.247	0.234	0.190
0.249	0.239	0.194
0.252	0.252	0.198
0.252	0.247	0.196
0.256	0.247	0.200
0.252	0.254	0.197
0.254	0.249	0.200
0.252	0.243	0.199
0.251	0.240	0.195
0.251	0.245	0.196

Table 4: *Maximum likelihood estimates for the full data set (first column), the mean of the ML estimates of 5 squashed data sets (second column) and 10 000 sub-sampled data sets (third column) for 15 independent data set of 100 000 samples from the GPD with* $\xi = 0.25$. *The squashed and sub-sampled data sets were found using 10 regions and 3 point per region. The top and bottom lines are the result of the full experiment of Tables 2 and 3 and the mean over the 15 independent data sets respectively.*

closely and the surface (3) searched extensively to obtain more information on is shape. This could involve allowing more iterations, adjusting the optimization weights or applying other optimization routines.

Our simulation experiment is limited to two test distributions only. We believe that it is of interest to do a full study of data squashing for GPDs with even heavier tails. As indicated above, some preliminary results suggest that data squashing could be more difficult for distributions with infinite expectations.

The effect of the grouping is an issue of general importance for data squashing. Our results do not indicate that any particular grouping strategy is preferable, but we believe that an experiment in which the the number of regions, the size of the tail region and the number of squashed point in each region is varied beyond the study of this paper, is of interest.

Our study has been restricted to using maximum likelihood estimation to compare and evaluate data squashing and sub-sampling. Other possible criteria for comparison are the method of probability-weighted moments or direct comparison of estimated quantiles, see Embrechts et al. (1997, Section 6.3.2).

5 Conclusions

This paper has presented the general technique of data squashing for reducing a massive data set and applied it to data from the generalized Pareto distribution. The generalized Pareto distribution was chosen because it provides a range of heavy tailed distributions with finite as well as infinite moments. A natural competitor to data squashing is sub-sampling techniques, and it has been our purpose to compare the two techniques with respect to maximum likelihood estimation.

Our simulation results show that data squashing can be very useful for heavy tailed data. For our test distributions data squashing clearly captures the likelihood better than what is achieved by stratified random sampling and estimation is associated with a substantially smaller variance. The best results are obtained for the test distribution with the least heavy tail.

The test data sets that have been squashed and sub-sampled are small and their size presents no obstacle to any statistical method. However, the results are interesting as they indicate that data squashing can improve on sub-sampling for higher dimensional and larger data sets with heavy tail characteristics. Although there are aspects of

our results that we are unable to explain fully and there is a need for further work to understand the properties of data squashing for heavy tailed data, we feel that the results are promising.

Acknowledgments

This work was supported by The Research Council of Norway (NFR) under grant no. 110673/420 (Numerical Computations in Applied Mathematics). The software used for the simulation experiments was developed by Ragnar B. Huseby at the Norwegian Computing Center under grant no. 121144/420 (Knowledge, data and decisions). The simulation experiments were conducted using computing facilities at the Norwegian Computing Center. The author wishes to thank Ragnar B. Huseby and Erlend Berg for introducing her to data squashing and for many interesting discussions. Many thanks to Ola Haug for guidance with the S-Plus software of Alexander McNeil and to Arnoldo Frigessi for many useful comments.

References

Berg, E., Dimakos, X. K., and Huseby, R. B. (2000). Squashing massive data sets: An overview of existing methods and ideas for further research. Norwegian Computing Center, Report no. 961.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999). Squashing flat files flatter. In *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining*, pp. 6–15.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events*. Springer-Verlag, Berlin Heidelberg.

Johnson, T. and Dasu, T. (1998). Comparing massive high dimensional data sets. In *Proc. of the 4h Intl. Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 229–233.

Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., and Ridgeway, G. (2000). Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*. To appear.

McNeil, A. (2000). http://www.math.ethz.ch/~mcneil/software.html. EVIS Software for Extreme Values in S-Plus.

Owen, A. (1999). Data squashing by empirical likelihood. Available at http://www-stat.stanford.edu/~owen/reports/.