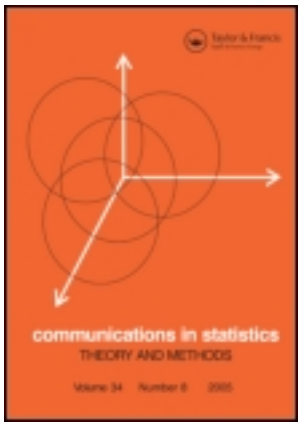

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Precision Index in the Multivariate Context

Miroslav Šiman [a]
[a] Institute of Information Theory and Automation of the ASCR , Prague , Czech Republic
Published online: 27 Dec 2014.

Communications in Statistics—Theory and Methods, 43: 377–387, 2014

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2012.661509

Precision Index in the Multivariate Context

MIROSLAV ŠIMAN

Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

General multivariate quantiles are employed to extend the classic univariate process precision index to the multivariate context under very mild conditions. Using halfspace depth regions for this purpose is especially recommended because it leads to both computational simplicity and natural generalizations to the tool-wear setup thanks to some recent advances in multiple-output and projectional quantile regression. A few examples are included to illustrate how the methodology might work in practice.

Keywords Data depth; Multivariate quantile; Precision index; Process capability index; Regression quantile.

Mathematics Subject Classification Primary 62P30; Secondary 62G05, 62G15.

1. Introduction

Process capability indices (PCIs) were introduced for quantifying the acceptability of manufacturing processes and soon became indispensable for any quality improvement program.

The theory of PCIs is already rich and quite developed in the univariate case where there is only one important stochastic feature $Y \in \mathbb{R}$ of the output. Typically, we have an idea about the optimal target value $T \in \mathbb{R}$ of the process as well as about the range $\mathcal{T} = [LSL, USL]$ of all conforming output values, and the goal of PCIs then consists in quantifying how much Y meets the ideal. For example, the potential capability of the process can be assessed by the popular precision index C_p

$$C_p = \frac{USL - LSL}{6\sigma}$$

where σ stands for a scatter parameter of Y, typically $\sigma = \sqrt{\mathrm{var}(Y)}$. This important index is also a cornerstone on which more sophisticated indices are built, see e.g.,

Received September 26, 2011; Accepted January 24, 2012

Address correspondence to Miroslav Šiman, Institute of Information Theory and Automation of the ASCR, Pod Vodárenskou věží 4, CZ-182 08, Prague 8, Czech Republic; E-mail: siman@utia.cas.cz

Pearn and Kotz (2006), and this is why its natural generalization to the multivariate setup is highly desirable and in the center of our current interest.

In the multivariate context, we have to deal with a stochastic vector process characteristic $\mathbf{Y} \in \mathbb{R}^m$, with its optimal target value $\mathbf{T} \in \mathbb{R}^m$ and with the tolerance region $\mathcal{T} \subset \mathbb{R}^m$ of all conforming values of $\mathbf{Y}$. Current proposals of multivariate precision indices are briefly summarized in Pearn and Kotz (2006) and Kotz and Lovelace (1998). Basically, these suggestions often assume $\mathcal{T}$ in a special shape and the distribution of $\mathbf{Y}$ normal or elliptical, employ variance matrices and related principal vectors or Mahalanobis distances, use elliptical quantile regions, and consider their volumes or their largest inflated copies still contained in the tolerance region. And if some ellipsoids are considered, they are often centered around $\mathbf{T}$ or a location parameter $\boldsymbol{\mu}$ of the distribution of $\mathbf{Y}$.

Of all these ideas, we loosely employ only that of inflation, see, e.g., Chen (1994) and Taam et al. (1993), but without any information about $\mathbf{T}$ or $\boldsymbol{\mu}$ required. Contrary to the prevailing practice, we are now going to introduce a nonparametric multivariate precision index without any limiting assumption on $\mathcal{T}$ and the distribution of $\mathbf{Y}$, i.e. without any restrictive condition on its moments, location, or quantile regions. This is possible only because we build our definition on a general nonparametric multivariate quantile concept.

Although our proposal could be based on any notion of nonparametric multivariate quantile mentioned in the literature, see, e.g., Serfling (2002) and references therein, we formulate our definition only for multivariate quantile regions defined as halfspace depth regions because they beat the others in many respects and have many favorable properties that simplify their application and computation considerably, see, e.g., Rousseeuw and Ruts (1999) and Zuo and Serfling (2000a,b). For example, they are always closed, convex, compact, and fully affine equivariant. In the empirical case, they are moreover polyhedral and thus easy to handle and compute, and all of them together fully determine the underlying empirical distribution. In the population case, this equivalence between the knowledge of distribution and of halfspace depth regions has been fully established only when all the regions have smooth boundaries, but what is more important, the population halfspace depth regions are always elliptical for elliptical distributions and the deepest one always coincides with the center of symmetry if such a point exists, see also Struyf and Rousseeuw (2005), Kong and Mizera (2012), and Kong and Zuo (2010). Consequently, these halfspace depth regions can be viewed as a natural extension of elliptical quantile regions to the general multivariate context.

If the manufacturing process is stable (stationary), then we can easily obtain a random sample from the distribution of $\mathbf{Y}$ and use it for sophisticated statistical inference, especially if we have a concrete idea about the distribution of $\mathbf{Y}$. This distribution is often automatically assumed perfectly normal after some questionable invocations to simplicity and central limit theorems so as the fraction of nonconforming items $P(\mathbf{Y} \notin \mathcal{T})$ could be estimated easily, which is a quantity of paramount importance in the quality management context. Unfortunately, the reality tends to be far more complicated and, what is even worse, $P(\mathbf{Y} \notin \mathcal{T})$ is usually highly sensitive even to changes in the distribution of $\mathbf{Y}$ too small to be detected reliably by any statistical test even from quite large random samples. This is why we do not feel bound by the normality dogma any more and formulate our concept without any distributional assumption at all, at the expense of obscuring $P(\mathbf{Y} \notin \mathcal{T})$. Nevertheless, if a specific and reliable assumption about the distribution

of $\mathbf{Y}$ is available, then $P(\mathbf{Y} \notin \mathcal{T})$ might still be estimated by means of Monte Carlo simulations. That is to say that current computers can easily generate billions of random vectors in a moment or two.

This paper introduces basic definitions and notation in Section 2, presents a multivariate precision index in Section 3, illustrates the concept in Section 4, and concludes with some further proposals, possible extensions, and final remarks in Section 5.

2. Definitions and Notation

Let us consider a multivariate product or process described by a continuous random vector $\mathbf{Y} \in \mathbb{R}^m$, $m \geq 1$, and assume that its engineering specification is always given by a convex and compact tolerance region $\mathcal{T} \subset \mathbb{R}^m$ of all conforming values, and sometimes also by a target value $\mathbf{T} \in \mathcal{T}$ in the interior of $\mathcal{T}$. The convexity assumption on $\mathcal{T}$ still allows for all the common tolerance regions of elliptical and polyhedral shapes, and it guarantees that the segment linking any $\mathbf{T} \in \mathcal{T}$ with any $\mathbf{y} \in \mathcal{T}$ contains only conforming points, which is usually a highly desirable property. The compactness of $\mathcal{T}$ is not limiting at all because everything real and material is bounded anyway and $\partial\mathcal{T}$ is too negligible to have any effect.

If the target value $\mathbf{T} \in \mathcal{T}$ is also provided, we will always assume it zero without any loss of generality as we can always shift the coordinate system accordingly.

We will also write $\mathcal{Q}_v(\cdot)$ for the multivariate halfspace depth region of the argument that is indexed by $v \in \mathbb{R}$ where v may stand either for the coverage probability $p \in (0, 1)$ or for the probability mass $\tau \in (0, 0.5]$ cut off by supporting hyperplanes. The corresponding halfspace depth median of the argument will be denoted by $\mathbf{m}(\cdot) \in \mathbb{R}^m$, and the sample counterparts to $\mathcal{Q}_v(\cdot)$ and $\mathbf{m}(\cdot)$ will be denoted by $\widehat{\mathcal{Q}}_v(\cdot)$ and $\widehat{\mathbf{m}}(\cdot)$, respectively. The argument will be omitted if it clearly follows from the context or if we speak about halfspace depth regions or medians in general.

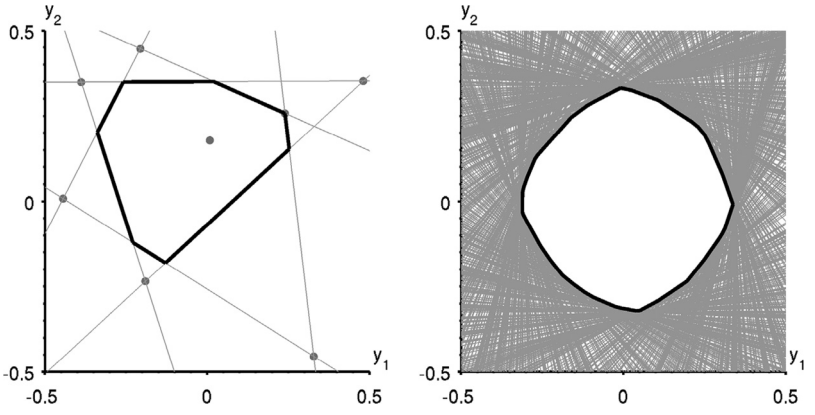

Figure 1. These two plots show empirical halfspace depth regions (*with the thick black border*) for $\tau = 1/6$ that are constructed from $n = 8$ visible dark gray points (*left*) or from $n = 1\,000$ invisible points coming from the uniform distribution on $[-0.5, 0.5]^2$ (*right*). The light gray lines (passing through at least $m = 2$ points) border all the closed halfspaces containing $n - k + 1$ data points where $k = \lfloor n\tau \rfloor + 1$.

Recall that $\mathcal{Q}_\tau(\mathbf{Y})$ is defined as the intersection of all closed halfspaces H with $P(\mathbf{Y} \in H) \geq 1 - \tau$ and that this region is always non empty at least for any $\tau \leq 1/(m+1)$, see Rousseeuw and Ruts (1999). In the empirical case with $n \geq m+1$ observations and with a positive $\tau \in [(k-1)/n, k/n)$ for some integer $k \leq n$, the empirical halfspace depth region $\widehat{\mathcal{Q}}_\tau(\mathbf{Y})$ can equivalently be obtained as the intersection of all closed halfspaces (a) containing at least $n-k+1$ observations, (b) containing exactly $n-k+1$ observations, or, if the interior of $\widehat{\mathcal{Q}}_\tau(\mathbf{Y})$ is non-empty, (c) containing exactly $n-k+1$ observations with at least m of them in the boundary, see Donoho and Gasko (1992) and Fukuda and Rosta (2005). Halfspace depth median is then usually defined as a specific point in the deepest non empty halfspace depth region. The remaining properties of halfspace depth regions are summarized in Section 1, their construction is illustrated in Figure 1, and their computation is discussed at the end of Section 4. Note as well that $\widehat{\mathcal{Q}}_\tau(\mathbf{Y})$ is equal to the convex hull of all the data points for any positive $\tau < 1/n$.

3. Multivariate Extension

We believe that the true multivariate precision index C_P should be independent of the target value and of any parameter of location. Therefore, we propose to define it as follows:

$$C_P = \kappa_{P,v} s_{\max}, \quad s_{\max} = \sup\left\{s > 0 : \mathcal{Q}_v(s\mathbf{Y} + \mathbf{c}_s) \subset \mathcal{T} \text{ for some } \mathbf{c}_s \in \mathbb{R}^m\right\} \qquad (1)$$

where $\kappa_{P,v} > 0$ is a normalizing scale factor. This definition makes sense for both $v \equiv p$ and $v \equiv \tau$. Besides, each value of this parameter can give rise to a different definition since, in general, $\mathcal{Q}_v(s\mathbf{Y})$ need not have the same shape for different values of v any more. The constant $\kappa_{P,v}$ then may be used to calibrate the index to match C_p in the univariate case, for example. Of course, fully affine equivariance of $\mathcal{Q}_v$ guarantees $\mathcal{Q}_v(s\mathbf{Y} + \mathbf{c}) = s\mathcal{Q}_v(\mathbf{Y}) + \mathbf{c}$.

Note that $s_{\max}$ has a very natural interpretation. It says how many times we can uniformly increase the production error on condition that the multivariate halfspace depth region $\mathcal{Q}_v$ of the output characteristic still lies inside $\mathcal{T}$ after a suitable adjustment (shift). In other words, it really measures the potential capability of the multivariate process. Therefore, C_P might be used as a perfect single scalar characteristic for measuring the performance of a multivariate process and for comparing two processes.

In fact, the well-known univariate index C_p can also be interpreted in the same way, for it tells us (up to a scalar multiplicative constant) how much the (non-extreme) confidence/quantile interval of a pre-agreed width $c\sigma$, $c \in \mathbb{R}_+$, or of the corresponding (non extreme) level can be stretched while still remaining in the tolerance interval after a suitable shift.

It remains to describe how such an index could be computed in the empirical case. As any convex body can be well approximated by an intersection of a sufficiently large number of supporting halfspaces, we may always replace $\mathcal{T}$ with its accurate polyhedral approximation, say $\widehat{\mathcal{T}}$, without any significant loss of generality. Of course, this $\widehat{\mathcal{T}}$ can be described by coordinatewise vector inequalities, say $A\mathbf{y} \leq \mathbf{r}_A$, where $\mathbf{r}_A$ must have all its coordinates positive if $\mathbf{T} = \mathbf{0}$ lies in the interior of $\mathcal{T}$. Furthermore, all the empirical halfspace depth regions $\widehat{\mathcal{Q}}_v(s\mathbf{Y})$ are polyhedral by definition.

Let us define $V_s = (\mathbf{v}_s(1), \ldots, \mathbf{v}_s(N_V))$ as a matrix containing all the vertices $\mathbf{v}_s(i)$, $i = 1, \ldots, N_V$, of $\widehat{\mathcal{Q}}_v(s\mathbf{Y})$ in its columns, obviously $V_s = sV_1$. Then the set $\widehat{\mathcal{Q}}_v(s\mathbf{Y} + \mathbf{c}_s)$ lies in $\widehat{\mathcal{T}}$ for some $\mathbf{c}_s$ if and only if each of its vertices does, i.e. if and only if the system of linear inequalities $A(\mathbf{v}_s(1) + \mathbf{c}) \leq \mathbf{r}_A, \ldots, A(\mathbf{v}_s(N_V) + \mathbf{c}) \leq \mathbf{r}_A$ has a feasible solution $\mathbf{c}$, which can be checked for any specific s quickly and easily by means of linear programming. Note also that the system of linear inequalities can be simplified to $A\mathbf{c} \leq \mathbf{r}_A - s \max_{i=1,\ldots,N_V} A\mathbf{v}_1(i)$ where the maximum is considered coordinatewise.

Therefore, $s_{\max}$ can be easily determined at least approximately by considering s from a sufficiently dense uniform grid of cardinality K. In fact, one can speed up the search for $s_{\max}$ by employing the core idea of the binary search algorithm as $s_1 \leq s_{\max}$ implies $s_2 \leq s_{\max}$ for any $0 < s_2 \leq s_1$. Consequently, only some $\log_2(K)$ values of s have to be investigated for finding a good approximation to $s_{\max}$, see Section 4 for an illustration.

If all the conforming values of $\mathbf{Y}$ were equally good, then $\widehat{m}(s_{\max}\mathbf{Y} + \mathbf{c}_{s_{\max}}) = s_{\max}\widehat{m}(\mathbf{Y}) + \mathbf{c}_{s_{\max}}$ would probably be a natural candidate for the best target value $\mathbf{T}$ from this point of view. Of course, this observation is useful only if the target value $\mathbf{T}$ can be controlled and possibly adjusted.

4. Illustration

Let us consider an artificial example to illustrate the concept introduced in the previous section. Concretely, let us assume (irrelevant) $\mathbf{T} = \mathbf{0}$ and $\mathcal{T} \subset \mathbb{R}^2$ as the set of all $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$ meeting the following inequalities:

$$-4 \leq y_1 \leq 5, \quad -4 \leq y_2 \leq 5, \quad y_1 + y_2 \geq -4, \quad \text{and}$$
$$y_1^2 + y_2^2 \leq 25 \text{ if both } y_1 > 0 \quad \text{and} \quad y_2 > 0.$$

This set $\mathcal{T}$ is still convex but already more complicated than those usually employed. As it is not polyhedral, we will replace it with its fine polyhedral approximation $\widehat{\mathcal{T}}$ obtained by replacing the last constraint with the set of N_ϕ linear limitations

$$\cos\left(\frac{i\pi}{2(N_\phi + 1)}\right) y_1 + \sin\left(\frac{i\pi}{2(N_\phi + 1)}\right) y_2 \leq 5, \quad i = 1, \ldots, N_\phi,$$

say $N_\phi = 45$ here. Figure 2 shows (in black) the border lines of all the linear constraints defining $\widehat{\mathcal{T}}$.

We would like to convince the reader that our concept works well even in the standard case, i.e., when applied to a moderate number of data points drawn from a normal distribution. This is why we first considered $n = 199$ (*dark gray*) observations simulated from the multivariate normal distribution with the mean vector $\boldsymbol{\mu} = (2, -1)'$ and the variance matrix V with $\text{vech}(V) = (0.5, 0.25, 1)'$. For (arbitrarily chosen) $v \equiv \tau = 0.10$, we quickly obtained the (*black*) sample halfspace depth region $\widehat{\mathcal{Q}}_\tau$ and its (*light gray*) shifted and inflated version $s_{\max}\widehat{\mathcal{Q}}_\tau + \mathbf{c}_{s_{\max}}$ still inside $\widehat{\mathcal{T}}$, where $\mathbf{c}_{s_{\max}} = (-5.39, 3.17)'$ and $s_{\max} = 3.33$; see the left panel of Figure 2. We can also add that the original sample halfspace depth region has its volume equal to 3.19, 22 vertices, and 109 observations inside.

How did we find $s_{\max}$? We started with $s = 250$ and with the lower bound $S_{\min} = 0$ and upper bound $S_{\max} = 500$ on $s_{\max}$. Next we set $S_{\min} = s$ if s lead to

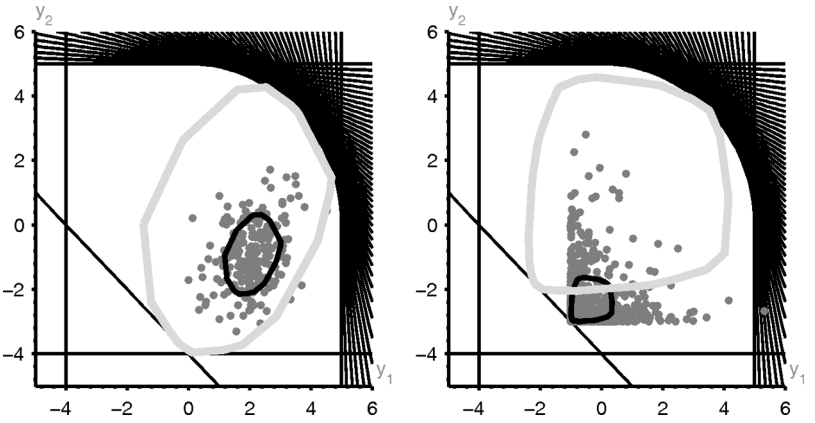

Figure 2. These two plots demonstrate how the new method works for $\tau = 0.1$, n (*dark gray*) observations and the tolerance region $\mathcal{T}$ approximately bordered by the straight black lines. The original sample halfspace depth region $\widehat{\mathcal{Q}}_\tau$ is circumscribed by the black curve while its maximal inflated copy still inside $\mathcal{T}$ after a suitable shift is enclosed by the light gray one. The left plot contains $n = 199$ data points generated from the multivariate normal distribution with the mean vector $\boldsymbol{\mu} = (2, -1)'$ and the variance matrix V, $\mathrm{vech}(V) = (0.5, 0.25, 1)'$, while the right plot displays $n = 999$ points simulated from the distribution of $(Y_1, Y_2)' = (-1, -3)' + (Z_1^2, Z_2^2)'/2$ where $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$ are independent.

a feasible solution and $S_{\max} = s$ otherwise, defined $s = (S_{\min} + S_{\max})/2$ for the next step, and repeated all that till the difference between two consecutive values of s was sufficiently small.

We also considered an example of a highly non normal bivariate distribution. With the same $\tau = 0.10$ as above, we generated $n = 999$ (dark gray) observations from the distribution of $(Y_1, Y_2)' = (-1, -3)' + (Z_1^2, Z_2^2)'/2$ where $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$ were independent. The resulting (black) original halfspace depth region and its (light gray) shifted and inflated version are plotted in the right panel of Figure 2. The latter region corresponds to $\mathbf{c}_{s_{\max}} = (2.46, 12.55)'$ and $s_{\max} = 4.88$. In this case, the original sample halfspace depth region has its volume equal to 1.55, 66 vertices, and 469 observations inside.

Finally, we include Figures 3 and 4 that display the results for fixed $n = 100$ observations coming from four different distributions and for a fixed normal distribution with the number of observations increasing from $n = 100$ to $n = 400$, respectively.

In all cases, we used Matlab 7.11 with SeDuMi 1.21 for necessary computations, see Sturm (1999) and Pólik (2005). Any concrete halfspace depth region can be determined simply from all the halfspaces with at least m points in the boundary because the number of available observations is usually quite small. Many sophisticated algorithms avoiding the need of all such halfspaces for computing a single halfspace depth region are also described in the literature, and some of them have already been implemented for bivariate data in Fortran, C++, C, R or Matlab; see Rafalin and Souvaine (2004) with references therein. The algorithms and Matlab codes presented in Paindaveine and Šiman (2012b,c) can even handle data beyond dimension two and in a general regression context.

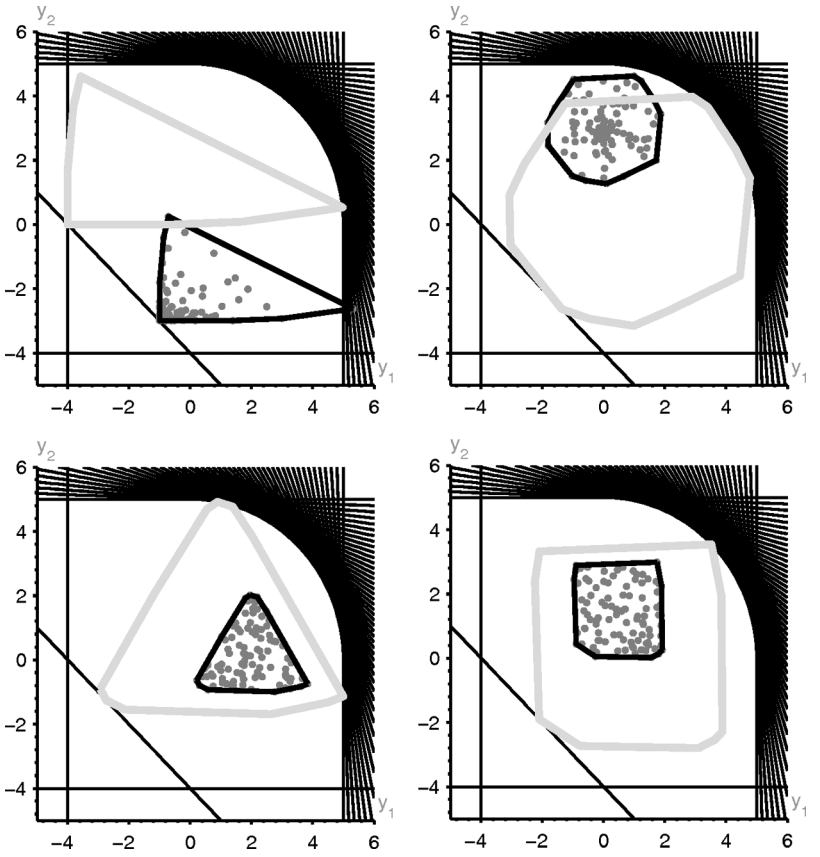

Figure 3. These plots show how the new method works for $n = 100$ (*dark gray*) data points, $\tau = 1/(2n)$, and the tolerance region $\mathcal{T}$ approximately bordered by the straight black lines. The original sample halfspace depth region (i.e., the convex hull of the observations) $\widehat{\mathcal{Q}}_\tau$ is circumscribed by the black curve while its maximal inflated copy still inside $\mathcal{T}$ after a suitable shift is enclosed by the light gray one. The independent observations $(Y_1, Y_2)'$ were simulated as follows: (a) $(Y_1, Y_2)' = (-1, -3)' + (Z_1^2, Z_2^2)'/2$ where $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$ are independent, (b) $(Y_1, Y_2)' = (0, 3)' + R(\cos(\phi), \sin(\phi))'$ where $\phi \sim U([0, 2\pi])$ and $R \sim U([0, 2])$ are independent, (c) $(Y_1, Y_2)'$ are uniformly distributed in the equilateral triangle with vertices $[0, -1]$, $[4, -1]$, and $[2, 2\sqrt{3} - 1]$, (d) $(Y_1, Y_2)'$ are uniformly distributed in the square with vertices $[-1, 0]$, $[2, 0]$, $[2, 3]$, and $[-1, 3]$.

We can conclude that the method usually leads to satisfactory results, even if the characteristic of the process is highly non-normal and the number of observations is small such as $n = 100$. Of course, larger numbers of data points can be expected to result in neater and less volatile outputs.

5. Final Remarks

Finally, we would like to point out several facts and bring up some ideas about what to do next.

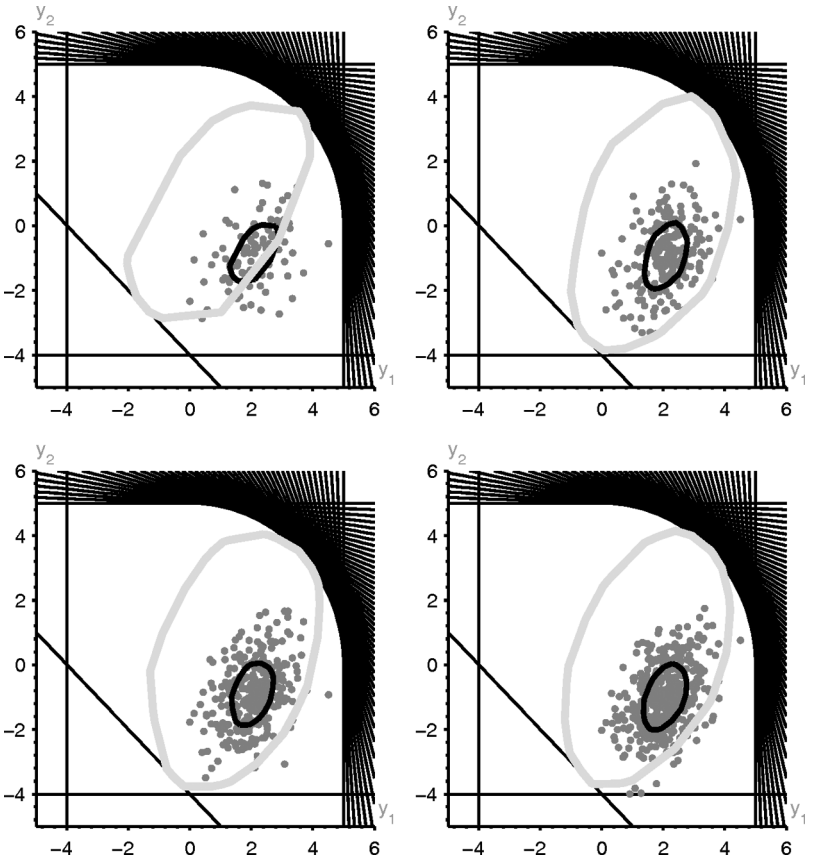

Figure 4. This figure demonstrates how the new method works for the tolerance region $\mathcal{T}$ approximately bordered by the straight black lines, $\tau = 0.158655$, and increasing number n of data points generated from the multivariate normal distribution with the mean vector $\boldsymbol{\mu} = (2, -1)'$ and the variance matrix V, $\mathrm{vech}(V) = (0.5, 0.25, 1)'$, where (a) $n = 100$ [$s_{\max} = 3.77$], (b) $n = 200$ [$s_{\max} = 3.83$], (c) $n = 300$ [$s_{\max} = 4.06$], and (d) $n = 400$ [$s_{\max} = 3.84$]. The original sample halfspace depth region $\widehat{\mathcal{Q}}_\tau$ is circumscribed by the black curve while its maximal inflated copy still inside $\mathcal{T}$ after a suitable shift is enclosed by the light gray one.

First, our method is not at all limited by the dimension m of $\mathbf{Y}$, at least theoretically. Unfortunately, the computation of sample halfspace depth regions becomes more time and space consuming with growing m, which currently makes this method impractical for $m \geq 4$ or 5, see Paindaveine and Šiman (2012b,c). In the previous section, we chose $m = 2$ only for the sake of clarity and simplicity, mainly because any plotting in spaces of dimensions higher than two is notoriously troublesome.

Second, this method can be easily generalized to the regression/tool-wear context. The only thing required is to replace halfspace depth regions with multivariate quantiles of the conditional distribution, that are again polyhedral in the empirical case. There are already some natural candidates for this post in the literature, including the regression cuts considered in Hallin et al. (2010) and Paindaveine and Šiman (2012a), and the local polynomial variants of halfspace

depth regions of Hallin et al. (accepted). In fact, the idea of using regression quantiles for such a purpose may be new even in the univariate case where it could be applied as well and based on the single-response regression quantiles, introduced in Koenker and Bassett (1978) and masterly reviewed in Koenker (2005).

Third, the technique presented here might also be successfully combined with the standard elliptical approach. That is to say that the standard elliptical quantile might be approximated by a polyhedron that could be subsequently inflated and shifted as described above, still without any use of $\mathbf{T}$ or $\boldsymbol{\mu}$. The same approach could be used even with any other concept of convex quantile regions. To be honest, there are also some quantile regions described in the literature that need not be necessarily convex, see, e.g., Hlubinka et al. (2010), but their convex hulls could still be employed in the same way.

Fourth, the value of v influences the robustness of the halfspace depth region as well as the overall computational time, see Paindaveine and Šiman (2012b,c). In fact, $v \equiv \tau$ can be easily employed as it is directly linked to the definition of halfspace depth regions. On the contrary, their coverage probability can be assessed only indirectly (and thus many halfspace depth regions would have to be computed to pick up the right one). Although the best choice of v should be left for some real experts to decide, we would definitely suggest to use $v \equiv \tau$.

Note that if the population distribution is elliptical or with identical independent symmetric stable marginals as in Chen and Tyler (2004), then the selection of τ is virtually irrelevant as it only scales the resulting coefficients C_P because all the halfspace depth regions are then of the same shape, i.e. only inflated copies of one another. Therefore the choice of τ is roughly irrelevant in the corresponding sample case as well. But it matters in general because the shapes of the halfspace depth regions may vary for different τ's.

In principle, there are a few possible approaches to choosing τ: (a) as a neat number as in Figure 2, (b) as an m-dependent neat number such as $\tau = 1/(2(m+1))$ in Figure 1, which ensures $\widehat{\mathcal{Q}}_\tau$ non empty in any dimension, (c) as a number with a direct link to the normal case such as $\tau = 0.158655$ in Figure 4, which leads to the coefficient C_P three times higher than C_p in the normal univariate case if $\kappa_{P,\tau} = 1$, (d) $\tau = 1/(2n)$ as in Figure 3, which produces empirical halfspace depth regions equal to the convex hulls of the data points. It would also be possible to consider several τ's at once. Clearly, each choice has its benefits and drawbacks. We find (d) especially appealing because it is very intuitive and avoids any sophisticated theory, but it is up to the practitioners to agree on the selection rule for τ that best fits their needs, just as they decided to use the 6σ methodology rather than the 7σ one.

Fifth, the method could be easily modified by scaling only the individual coordinates of $\mathbf{Y} = (Y_1, \ldots, Y_m)'$. Indeed, we might define m precision indices $C_P^i = \kappa_{P,v} s_{\max}^i$, $i = 1, \ldots, m$,

$$s_{\max}^i = \sup\left\{s > 0 : \mathcal{Q}_v((Y_1, \ldots, sY_i, \ldots, Y_m)' + \mathbf{c}_s^i) \subset \mathcal{T} \text{ for some } \mathbf{c}_s^i \in \mathbb{R}^m\right\} \quad (2)$$

and consider them together to assess the precision of a multivariate process. They could be interpreted and computed like C_P, thanks to the affine equivariance of halfspace depth regions.

Sixth, in the sample case, $s_{\max}$ and $s_{\max}^i$, $i = 1, \ldots, m$, become random variables whose quantiles or variances are clearly essential for evaluating the results. These quantities might be naturally estimated by a resampling procedure such as

jacknife or bootstrap, but the optimal construction, finite-sample performance, and asymptotic behavior of such estimators are yet to be investigated.

Seventh, those univariate PCIs using the target value should also be generalized to the multivariate context, and a natural way to do so is discussed in the accompanying paper Šiman (in press).

Eighth, all the PCIs mentioned in this paper could also be used quite naturally in control charts for monitoring the evolution of multivariate processes in the course of time.

Acknowledgments

The author would like to thank Jiří Michálek for careful reading of the manuscript and Davy Paindaveine, Marc Hallin, Claude Adan, Nancy de Munck, and Romy Genin for all the good they did for him (and for all the good he could learn from them) during his stay at Université Libre de Bruxelles.

Funding

This research work of Miroslav Šiman was supported by Project 1M06047 of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

Chen, H. F. (1994). A multivariate process capability index over a rectangular solid tolerance zone. *Statistica Sinica* 4:749–758.

Chen, Z., Tyler, D. E. (2004). On the behavior of Tukey's depth and median under symmetric stable distributions. *J. Statist. Plan. Infer.* 1–2:111–124.

Donoho, D. L., Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* 20:1803–1827.

Fukuda, K., Rosta, V. (2005). Data depth and maximum feasible subsystems. In: Avis, D., Hertz, A., Marcotte, O., eds. *Graph Theory and Combinatorial Optimization.* New York: Springer, pp. 37–67.

Hallin, M., Paindaveine, D., Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. *Ann. Statist.* 38:635–669.

Hallin, M., Lu, Z., Paindaveine, D., Šiman, M. Local constant and local bilinear multiple-output quantile regression. (accepted).

Hlubinka, D., Kotík, L., Vencálek, O. (2010). Weighted halfspace depth. *Kybernetika* 46:125–148.

Koenker, R., Bassett, G. J. (1978). Regression quantiles. *Econometrica* 46:33–50.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Kong, L., Mizera, I. (2012). Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica* 22:1589–1610.

Kong, L., Zuo, Y. (2010). Smooth depth contours characterize the underlying distribution. *J. Multivar. Anal.* 101:2222–2226.

Kotz, S., Lovelace, C. R. (1998). *Process Capability Indices in Theory and Practice*. New York: Oxford University Press.

Paindaveine, D., Šiman, M. (2012a). On directional multiple-output quantile regression. *J. Multivar. Anal.* 102:193–212.

Paindaveine, D., Šiman, M. (2012b). Computing multiple-output regression quantile regions. *Comp. Statist. Data Anal.* 56:840–853.

Paindaveine, D., Šiman, M. (2012c). Computing multiple-output regression quantile regions from projection quantiles. *Comp. Statist.* 27:29–49.

Pearn, W. L., Kotz, S. (2006). *Encyclopedia and Handbook of Process Capability Indices*. Singapore: World Scientific Publishing.

Pólik, I. (2005). Addendum to the SeDuMi user guide: version 1.1. reference guide.

Rafalin, E., Souvaine, D. L. (2004). Computational geometry and statistical depth measures. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S., eds. *Theory and Applications of Recent Robust Methods.* Basel: Birkhauser, pp. 283–296.

Rousseeuw, P. J., Ruts, I. (1999). The depth function of a population distribution. *Metrika* 49:213–244.

Serfling, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Statistica Neerlandica* 56:214–232.

Struyf, A., Rousseeuw, P. J. (2005). Halfspace depth and regression depth characterize the empirical distribution. *J. Multivar. Anal.* 69:135–153.

Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Opt. Meth. Software* 11–12:625–653.

Šiman, M. Multivariate capability indices: A directional approach. *Commun. Statist. Theor. Meth.* (in press).

Taam, W., Subbaiah, P., Liddy, J. W. (1993). A note on multivariate capability indices. *J. Appl. Statist.* 20:339–351.

Zuo, Y., Serfling, R. (2000a). General notions of statistical depth functions. *Ann. Statist.* 28:461–482.

Zuo, Y., Serfling, R. (2000b). On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *J. Statist. Plan. Infer.* 84:55–79.