Working Papers

02/2010

A Class of stochastic optimization problems with application to selective data editing

Ignacio Arbués

Margarita González

Pedro Revilla

First draft: December 2010*

This draft: December 2010

*This is a preprint of an article whose final and definitive form has been published in OPTIMIZATION © 2010 Taylor and Francis; OPTIMIZATION is available online at http://www.informaworld.com/smpp/content~db=all~content=a926591539~frm=titlelink

A Class of stochastic optimization problems with application to selective data editing

Abstract

We present a new class of stochastic optimization problems where the solutions belong to an infinite-dimensional space of random variables. We prove existence of the solutions and show that under convexity conditions, a duality method can be used. The search for a good selective editing strategy is stated as a problem in this class, setting the expected workload as the objective function to minimize and imposing quality constraints.

Keywords

Stochasting Programming, Optimization in Banach Spaces, Selective Editing, Score Function

Authors and Affiliations

Ignacio Arbués

Dirección General de Metodología, Calidad y Tecnologías de la Información y las Comunicaciones, Instituto Nacional de Estadística

Margarita González

D. G. de Coordinación Financiera con las Comunidades Autónomas y con las Entidades Locales, Ministerio de Economía y Hacienda

Pedro Revilla

Dirección General de Metodología, Calidad y Tecnologías de la Información y las Comunicaciones, Instituto Nacional de Estadística

A Class of stochastic optimization problems with application to selective data editing

Ignacio Arbués[†,*], Margarita González[‡], Pedro Revilla[†]

December 22, 2010

[†] *D. G. de Metodología, Calidad y Tecnologías de la Información y las Comunicaciones, Instituto Nacional de Estadística, Madrid, Spain*

[‡] *D. G. de Coordinación Financiera con las Comunidades Autónomas y con las Entidades Locales, Ministerio de Economía y Hacienda, Madrid, Spain*

Abstract

We present a class of stochastic optimization problems with constraints expressed in terms of expectation and with partial knowledge of the outcome in advance to the decision. The constraints imply that the problem cannot be reduced to a deterministic one. Since the knowledge of the outcome is relevant to the decision, it is necessary to seek the solution in a space of random variables. We prove that under convexity conditions, a duality method can be used to solve the problem. An application to statistical data editing is also presented. The search of a good selective editing strategy is stated as an optimization problem in which the objective is to minimize the expected workload with the constraint that the expected error of the aggregates computed with the edited data is below a certain constant. We present the results of real data experimentation and the comparison with a well known method.

Keywords: Stochastic Programming; Optimization in Banach Spaces; Selective Editing; Score Function

AMS Subject Classification: 90C15; 90C46; 90C90

1 Introduction

Let us consider the following elements: a decision variable x under our control; the outcome ω of a probability space; a function f that depends on x and ω that we want maximize; and a function g also depending on x and ω that we want in some sense not to exceed zero. With these materials, various optimization problems can be posed. In first place, the problem is quite different depending on whether we know ω in advance to the decision on x or not. In the first case

[*]Corresponding author. Email: iarbues@ine.es. Address: Instituto Nacional de Estadística, Castellana 183, 28071, Madrid, Spain.

(wait and see), when we make the decision ω is fixed and we have deterministic functions $f(\omega, \cdot)$ and $g(\omega, \cdot)$. Therefore, we can solve a nonstochastic optimization problem for each outcome ω, so that the solution will depend on ω and consequently to have a stochastic nature, so it would be convenient to study its properties as a random variable (for example, [2]). In the case ω is unknown (here and now), the usual approach is to pose a stochastic optimization problem by setting as the new objective function $\mathbf{E}[f(x, \omega)]$ (see [4] for a discussion on this matter). With respect to the function g, the most usual approach is to set the constraint $P[g(x, \omega) > 0] \leq p$ for a certain p.

The class of problems we study in this paper is an intermediate one with respect to the knowledge of ω, because we do not know it precisely, but we have some partial information about it, in a sense that will be specified in detail in the subsequent section. The use of this partial information to choose x implies that it will be a random variable itself. On the other hand, in our problems, the constraints are expressed in terms of expectation, that is, $\mathbf{E}[g(x, \omega)] \leq 0$.

We also analyse an application of these problems to statistical data editing. Efficient editing methods are critical for the statistical offices. In the past, it was customary to edit manually every questionnaire collected in a survey before computing aggregates. Nowadays, exhaustive manual editing is considered inefficient, since most of the editing work has no consequences at the aggregate level and can in fact even damage the quality of the data (see [1] and [5]).

Selective editing methods are strategies to select a subset of the questionnaires collected in a survey to be subject to extensive editing. A reason why this is convenient is that it is more likely to improve quality by editing some units than by editing some others, either because the first ones are more suspect to have an error or because the error if it exists has probably more impact in the aggregated data. Thus, it is reasonable to expect that a good selective editing strategy can be found that balances two aims: (i) good quality at the aggregate level and (ii) less manual editing work.

This task is often done by defining a score function (SF), which is used to prioritise some units. When several variables are collected for the same unit, different *local* score functions may be computed and then, combined into a *global* score function. Finally, those units with score over a certain threshold are manually edited.

Thus, when designing a selective editing process it is necessary to decide,

- Whether to use SF or not.
- The local score functions.
- How to combine them into a global score function (sum, maximum, ...).
- The threshold.

At this time, the points above are being dealt with in an empirical way because of to the lack of any theoretical support. In [7], [8] and [5] some guidelines are proposed to build score functions, but they rely in the criterion of the practitioner. In this paper, we describe a theoretical framework which, under some

assumptions, answers the questions above. For this purpose, we will formally define the concept of *selection strategy*. This allows to state the problem of selective editing as an optimization problem in which the objective is to minimize the expected workload with the constraint that the expected remaining error after editing the selected units is below a certain bound.

We present in section 2 the general problem. We also describe how to use a duality method to solve the problem. In section 3, we show the results of a simulation experiment. In sections 4 through 7 we describe how to apply the method to selective editing. In section 8, results of the application of the method to real data are presented. Finally, some conclusions are discussed.

2 The general problem

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $N, m, p \in \mathbb{N}$, let us consider the functions f defined from $\mathbb{R}^N \times \Omega$ in $\mathbb{R}$, $g_1 = (g_1^1, \ldots, g_1^m)$ defined from $\mathbb{R}^N$ in $\mathbb{R}^m$ and $g_2 = (g_2^1, \ldots, g_2^p)$ from $\mathbb{R}^N \times \Omega$ in $\mathbb{R}^p$. Let us also consider x a $N \times 1$ random vector. If θ is a function defined in $\mathbb{R}^N \times \Omega$, we will use the notation $\theta(x)$ for the mapping $\omega \in \Omega \mapsto \theta(x(\omega), \omega)$.

We consider the following problem,

$$[P] \quad \max \quad \mathbf{E}[f(x)] \tag{1}$$

$$\text{s.t.} \quad x \in M(\mathcal{G}), g_1(x) \leq 0 \ a.s., \mathbf{E}[g_2(x)] \leq 0. \tag{2}$$

where $M(\mathcal{G})$ is the set of the $\mathcal{G}-$measurable random variables, for a certain $\sigma-$field $\mathcal{G} \subset \mathcal{F}$. The condition $x \in M(\mathcal{G})$ is the formal expression of the idea of *partial information*. The necessity of seeking x among the $\mathcal{G}-$measurable random variables is the consequence of assuming that all we know about ω is whether $\omega \in A$, for any $A \in \mathcal{G}$, that is, if the event A happened.

In order to introduce the main features of the problem, we will analyse first the case that $\mathcal{G} = \mathcal{F}$ (full information) and then, we will describe how to reduce the general case (partial information) to the former one. In the next subsection, we will present conditions under which a duality method can be applied to solve problem $[P]$ with $\mathcal{G} = \mathcal{F}$.

2.1 Duality in the case of full information

Let us make some assumptions,

Assumption 1. *f and g_2 are measurable in ω.*

Assumption 2. *$-f$, g_1 and g_2 are convex in x and the function $x \in L^\infty(\Omega) \mapsto \mathbf{E}[\theta(x)]$ is lower semicontinuous for $\theta = -f, g_2$.*

Assumption 3. *There is a random vector x_0 such that $g_1(x_0) \leq 0$ a.s. and $\mathbf{E}[g_2(x_0)] < 0$.*

Assumption 4. *The set $\{z \in \mathbb{R}^N : g_1(z) \leq 0\}$ is bounded.*

Assumption 5. *$\mathcal{G}$ is countably generated.*

We will see that with assumptions 1–5, problem [P] is well posed. Assumptions 1, 2 are not too restrictive and then, f, g_1 and g_2 are allowed to range over a quite general class. Assumption 2 imply that f and g_2 are continuous in x and thus, they are Carathéodory functions. We can prove that if x is $\mathcal{G}-$measurable, then $\omega \mapsto \theta(x(\omega), \omega)$ is also $\mathcal{G}-$measurable and $\mathbf{E}[\theta(x)]$ makes sense. The assumptions on g_1 imply that the constraint $g_1(x) \leq 0$ *a.s.* is well defined. Assumption 3 is a classical regularity condition necessary for the duality methods and it is usually known as Slater's condition. Assumption 4 is restrictive and it is likely not necessary, but makes the proofs easier and it holds in our applications. Finally, 5 is a technical assumption that does not seem to imply an important loss of generality for most of practical applications.

We will solve $[P]$ by duality. Let us define the Lagrange function,

$$\mathcal{L}(x, \lambda) = \mathbf{E}[f(x)] - \lambda^T \mathbf{E}[g_2(x)]. \tag{3}$$

We can now define the problem,

$$\begin{array}{lll} [P(\lambda)] & \max_x & \mathcal{L}(\lambda, x) \\ & \text{s.t.} & g_1(x) \leq 0 \ a.s. \end{array}$$

The dual function is defined as $\varphi(\lambda) = \sup\{\mathcal{L}(\lambda, x) : g_1(x) \leq 0 \ \ a.s.\}$. If the supremum is a maximum, we denote by x_λ the point where it is attained. The dual problem is,

$$\begin{array}{lll} [D] & \min_\lambda & \varphi(\lambda) \\ & \text{s.t.} & \lambda \geq 0. \end{array}$$

This problem is of great interest for us because of the following proposition.

Proposition 1. *If assumptions 1–5 hold then,*

i) There exist solutions to $[P]$ and $[D]$.

ii) If x is a solution to $[P]$ and $\bar{\lambda}$ is a solution to $[D]$ then, x is a solution to $[P(\bar{\lambda})]$.

Proof. Let us see that the primal problem has a solution. From assumption 4, it follows that we can seek a solution in $E = L^\infty(\Omega)$, which is a Banach space (Theorem 3.11 in [13]). Under assumption 5, the closed unit ball B in E is weakly compact (see [3], p.246). From assumption 2 the set $M = \{x \in E : g_1(x) \leq 0 \ a.s., \mathbf{E}[g_2(x)] \leq 0\}$ is closed and convex and then, weakly closed. From assumption 4, it is also bounded. Then, there exists some $\varepsilon > 0$ such that $\varepsilon M \subset B$. Since εM is a weakly closed subset of a weakly compact set, it is weakly compact itself. M is also weakly compact because is homothetic to εM. On the other hand, $x \mapsto -\mathbf{E}[f(x)]$ is convex and lower semicontinuous. Thus, it is weakly lower semicontinuous and it attains a minimum in M. The existence of the solution to $[D]$ and *ii)* are granted by theorem 1, p. 224 in [9]. □

We will see how to compute x_λ and $\varphi(\lambda)$. The dual problem $[D]$ can be solved by numerical methods since it is a finite–dimensional optimization problem.

The advantage of $P[\lambda]$ is that the stochastic term is now in the objective function. Consequently, the solution can be easily obtained by solving a deterministic optimization problem for any outcome $\omega \in \Omega$,

$$\begin{array}{lll} [P_D(\lambda,\omega)] & \max_x & L(\lambda,x,\omega) \\ & \text{s.t.} & g_1(x) \leq 0 \end{array}$$

where $L(\lambda, x, \omega) = f(x,\omega) - \lambda^T g_2(x,\omega)$. This problem is related to $[P(\lambda)]$ by virtue of the following proposition.

Proposition 2. *There exists a solution x_λ to $[P(\lambda)]$ such that for any $\omega \in \Omega$,*

i) $x_\lambda(\omega)$ is a solution to $P_D(\lambda,\omega)$.

ii) $\varphi(\lambda) = \mathbf{E}[L(\lambda, x_\lambda(\omega), \omega)]$.

Proof. From assumptions, f and g_2 are Carathéodory functions and then, so is L. Therefore, L is random lower semicontinuous. This means that $\omega \to \operatorname{epi} L(\lambda,\cdot,\omega)$ is closed valued and $\mathcal{G}$-measurable (see [11]). Theorem 14.37 in [12] states that for a random lower semicontinuous function $G(x,\omega)$, the multivalued map $\Phi(\omega) = \arg\min\{G(x,\omega) : x \in \mathbb{R}^N\}$ is measurable. Since $C = \{x \in \mathbb{R}^N : g_1(x) \leq 0\}$ is convex and closed, it is easy to prove that $\Psi(\omega) = \arg\min\{L(\lambda,x,\omega) : x \in C\}$ is also measurable. Theorem 14.5 and corollary 14.6, again from [12], imply that there exists a measurable selection from Ψ, that is, a measurable function $x_\lambda(\omega)$ such that for any ω, $x_\lambda(\omega) \in \Psi(\omega)$.

Let y be a measurable function defined on Ω. Since for any ω, $x_\lambda(\omega)$ is a solution to $[P(\lambda,\omega)]$ then, $L(\lambda, y(\omega), \omega) \leq L(\lambda, x_\lambda(\omega), \omega)$. Taking expectation in both sides of the inequality, we get $\mathcal{L}(\lambda, y) \leq \mathcal{L}(\lambda, x_\lambda)$. Finally, since x_λ is a solution to $[P(\lambda)]$, $\varphi(\lambda) = \mathcal{L}(\lambda, x_\lambda) = \mathbf{E}[L(\lambda, x_\lambda)]$. □

The optimal $\bar{\lambda}$ will be obtained maximizing φ. Since φ is described in terms of expectation, it is necessary either to know the real distribution of the terms in L and compute explicitly its expectation or to estimate it. In practical applications, the explicit computation will usually not be feasible.

2.2 Partial information

Let us consider now the general case $\mathcal{G} \subset \mathcal{F}$. We will reduce the problem $[P]$ to,

$$\begin{array}{lll} [P^*] & \max_{x \in M(\mathcal{G})} & \mathbf{E}[f^*(x)] \qquad (4) \\ & \text{s.t.} & g_1(x) \leq 0 \ a.s., \mathbf{E}[g_2^*(x)] \leq 0. \qquad (5) \end{array}$$

where for any $z \in \mathbb{R}^N$, and $\omega \in \Omega$ we define $f^*(z,\omega) = \mathbf{E}[f(z,\cdot)|\mathcal{G}]$ and $g_2^*(z,\omega) = \mathbf{E}[g_2(z,\cdot)|\mathcal{G}]$. Both f^* and g^* are $\mathcal{G}-$measurable as functions of ω, so we can apply the results of the previous subsection with $\mathcal{G}$ instead of $\mathcal{F}$.

Of course, we have to prove first that the problem defined above is equivalent to $[P]$. In order to do this, we need the following generalization of the monotone convergence theorem.

Lemma 1. *Let $\{\xi_n\}_n$ be a sequence of random variables and $\mathcal{G}$ a $\sigma-$field. If $\xi_n(\omega) \nearrow \xi(\omega)$, then, $\mathbf{E}[\xi_n|\mathcal{G}] \nearrow \mathbf{E}[\xi|\mathcal{G}]$.*

Proof. The sequence of random variables $\eta_n := \mathbf{E}[\xi_n|\mathcal{G}]$ is nondecreasing. Consequently, $\eta_n(\omega) \nearrow \eta(\omega)$. We have only to check that $\eta = \mathbf{E}[\xi|\mathcal{G}]$. The random variable η is $\mathcal{G}-$measurable since it is limit of $\mathcal{G}-$measurable r.v.'s. On the other hand, for any $A \in \mathcal{G}$ it holds that,

$$\int_A \eta P(\mathrm{d}\omega) = \lim_n \int_A \mathbf{E}[\xi_n|\mathcal{G}]P(\mathrm{d}\omega) = \lim_n \int_A \xi_n P(\mathrm{d}\omega) = \int_A \xi P(\mathrm{d}\omega)$$

where the first and third identities hold by the monotone convergence theorem and the second by the definition of the conditional expectation (page 445 in [3]). □

With this lemma, we can prove the following proposition.

Proposition 3. *If $-f$, and g_2 are lower semicontinuous in x, then problem $[P^*]$ is equivalent to $[P]$.*

Proof. We only have to prove that $\mathbf{E}[\theta^*(x)] = \mathbf{E}[\theta(x)]$, for $\theta = -f, g_2$. Let us consider for $k = -n2^n, \ldots, n2^n - 1$, the intervals $I_{nk} = [k2^{-n}, (k+1)2^{-n})$ for $k = -n2^n - 1$, we set $I_{nk} = (-\infty, -n)$ and for $k = n2^n$, $I_{nk} = [n, +\infty)$. Then, we can define the functions $\theta_{n,k}(\omega) = \inf_{z \in I_{nk}} \theta(z, \omega)$ and,

$$\Psi_{nk}(z) = \begin{cases} 1 & \text{if} \quad z \in I_{nk} \\ 0 & \text{if} \quad z \notin I_{nk} \end{cases}$$

We can write,

$$\theta_n(z, \omega) = \sum_{k=-n2^n-1}^{n2^n} \theta_{n,k}(\omega)\Psi_{nk}(z).$$

Let us see that $\theta_n(z, \omega) \nearrow \theta(z, \omega)$. For any z and n, there exist k, l such that $z \in I_{n+1,k} \subset I_{n,l}$. Then, $\theta_n(z, \omega) \leq \theta_{n+1}(z, \omega) \leq \theta(z, \omega)$. Now, for any $\epsilon > 0$, there exists n_0 such that for any $n \geq n_0$, $|z - z'| \leq 2^{-n}$ implies $\theta(z', \omega) \geq \theta(z, \omega) - \epsilon$. Consequently, for any $n \geq n_0$, $\theta(z, \omega) \geq \theta_n(z, \omega) \geq \theta(z, \omega) - \epsilon$. Therefore, $\theta_n(z, \omega) \nearrow \theta(z, \omega)$.

Now, by lemma 1

$$\theta^*(z, \omega) = \mathbf{E}[\theta(z, \cdot)|\mathcal{G}] = \lim_n \sum_{k=-n2^n-1}^{n2^n} \Psi_{nk}(z)\mathbf{E}[\theta_{n,k}|\mathcal{G}].$$

If $x(\omega)$ is $\mathcal{G}-$measurable, then

$$\theta^*(x(\omega), \omega) = \lim_n \sum_{k=-n2^n-1}^{n2^n} \Psi_{nk}(x(\omega))\mathbf{E}[\theta_{n,k}|\mathcal{G}] = \mathbf{E}[\theta(x)|\mathcal{G}]$$

where we have used again the notation $\theta(x)$ for the random variable $\omega \mapsto \theta(x(\omega), \omega)$. Consequently, $\mathbf{E}[\theta^*(x)] = \mathbf{E}[\theta(x)]$. □

It is easy to see that $[P^*]$ satisfies assumptions 1 – 5. Since the functions involved in problem $[P^*]$ are measurable with respect to $\mathcal{G}$, it can be considered as a *full information* one and thus, solved by using the results of the previous subsection.

3 Simulation.

Proposition 1, provides the optimal solution to $[P]$, but the dual problem has to be solved by numerical methods in order to obtain the Lagrange multipliers. Thus, we have designed an example of problem $[P]$ such that λ can be computed analytically. On the other hand, we have obtained estimate values from simulation in order to compare with the true ones.

Let us consider the following case,

$$f(x) = \mathbf{1}^T x; \qquad g_1(x) = (x^T - \mathbf{1}^T, -x^T)^T; \qquad g_2(x, \omega) = \sum_{i=1}^{N} \delta_i(\omega) x_i - d,$$

where $\{\delta_i\}_{i=1,\ldots,N}$ are uniformly distributed in $[0, 1]$ and independent. It is easy to see that the solution to $[P(\lambda)]$ is,

$$x_i = \begin{cases} 1 & \text{if} \quad \lambda\delta_i < 1 \\ 0 & \text{if} \quad \lambda\delta_i > 1 \end{cases} . \tag{6}$$

Then, we can see that $\mathbf{E}[x_i] = \lambda^{-1}$ and $\mathbf{E}[x_i\delta_i] = (2\lambda^2)^{-1}$. Hence,

$$\varphi(\lambda) = \begin{cases} N(1 - \frac{\lambda}{2}) + \lambda d & \text{if} \quad \lambda < 1 \\ \frac{N}{2\lambda} + \lambda d & \text{if} \quad \lambda \geq 1 \end{cases} ,$$

and consequently, for $d < N/2$ the minimum is attained at $\bar{\lambda} = (\frac{N}{2d})^{1/2}$.

We will estimate $\bar{\lambda}$ by applying the sample–path optimization or sample average approximation method (see [10], [14]). We simulate a sample of size M of $\delta = (\delta^1, \ldots, \delta^M)$ and then we minimize the function $\hat{\varphi}(\lambda) = M^{-1} \sum_{j=1}^{M} \varphi_j(\lambda)$, where $\varphi_j(\lambda) = L(\lambda, x^j)$, $L(\lambda, x) = \mathbf{1}^T x - \lambda(\sum \delta_i x_i - d)$, $x^j = (x_1^j, \ldots, x_N^j)$ and x_i^j is defined as in (6) for the j-th simulated value of δ. The minimization of $\hat{\varphi}$ has been performed using the function *fmincon* of the mathematical pack MATLAB.

The results for a range of values of N and M are presented in table 1, suggesting that when N is large, even moderate values of M allow to achieve considerable accuracy. We have chosen $d = N/4$ and thus, the theoretical $\bar{\lambda}$ is $\sqrt{2}$.

N	M=1	M=5	M=10	M=25	M=50	M=100	M=250
10	0.2052	0.0966	0.0686	0.0413	0.0330	0.0236	0.0132
50	0.0919	0.0375	0.0313	0.0190	0.0134	0.0104	0.0057
100	0.0694	0.0278	0.0224	0.0117	0.0097	0.0062	0.0044
500	0.0286	0.0125	0.0097	0.0060	0.0044	0.0030	0.0018
1,000	0.0177	0.0090	0.0063	0.0047	0.0030	0.0018	0.0014
5,000	0.0089	0.0041	0.0032	0.0019	0.0014	0.0010	0.0006
10,000	0.0070	0.0028	0.0024	0.0012	0.0011	0.0007	0.0004

Table 1: RMS error in the estimation of $\bar{\lambda}$.

4 The selective editing problem

Let us introduce some notation,

- x_t^{ij} is the *true* value of variable j in questionnaire i at period t, with $i = 1, \ldots, N$ and $j = 1, \ldots, q$.
- $\tilde{x}_t^{ij} = x_t^{ij} + \varepsilon_t^{ij}$ is the *observed* value of variable j in questionnaire i at period t, ε_t^{ij} being the observation error.
- $X_t^k = \sum \omega_{ij}^k x_t^{ij}$ is the k-th statistic computed with the true values ($\tilde{X}_t^k$ is computed with the observed ones), with k ranging from 1 to p.

The linearity assumption implies a loss of generality, which is nevertheless not very important in the usual practice of statistical offices. Many statistics are in fact linear aggregates of the data, while in some other cases such as indices, they are ratios whose denominator depends on past values that can be considered as constant when editing current values. When the statistic is nonlinear, the applicability of the method will rely on the accuracy of a first-order Taylor expansion in $\{x_t^{ij}\}$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We assume that x_t^{ij} and ε_t^{ij} are random variables with respect to that space. There can be other random variables relevant to the selection process. Among them, some are known at the moment of the selection, such as $\tilde{x}_t^{ij}$, x_s^{ij} with $s < t$ or even variables from other surveys. The assumption that x_s^{ij} is known is equivalent to assume that when editing period t, the data from previous periods have been edited enough and does not contain errors. Deterministic variables such as working days may also be useful to detect anomalous data. We will denote by $\mathcal{G}_t$ the $\sigma-$field generated by all the information available up to time t. In order to avoid heavy notation, we omit the subscript t when no ambiguity arises.

Our aim is to find an adequate selection strategy. A selective editing strategy should indicate for any i whether questionnaire i will be edited or not and this has to be decided using the information available. In fact, we will allow the strategy not to determine precisely whether the unit is edited but only with a certain probability.

Definition 1. *A selection strategy (SS) with respect to $\mathcal{G}_t$ is a $\mathcal{G}_t$-measurable random vector $r = (r_1, \ldots, r_N)^T$ such that $r_i \in [0, 1]$.*

We denote by $S(\mathcal{G}_t)$ the set of all the SS with respect to $\mathcal{G}_t$. The interpretation of r is that questionnaire i is edited with probability $1 - r_i$. To allow $0 \leq r_i \leq 1$ instead of the more restrictive $r_i \in \{0, 1\}$ is theoretically and practically convenient because then, the set of strategies is convex and techniques from convex optimization can be used. Moreover, it could happen that the optimal value over this generalized space were better than the restricted case (just as in hypothesis testing a randomized test can have a greater power than any nonrandomized one). If for a certain unit, $r_i \in (0, 1)$, then the unit is effectively edited depending on whether $\chi_t^i < r_i$, where χ_t^i is a random variable distributed uniformly in the interval $[0, 1]$, and independent from every other variable in our framework (in order to accommodate these further random variables, we consider an augmented probability space, $(\Omega^*, \mathcal{F}^*, P^*)$, which is the product space of the original one times the suitable choice of $(\Omega_1, \mathcal{F}_1, P_1)$; only occasionally we have to refer to the augmented one). We denote by $\tilde{r}_i$ the indicator variable of the event $\chi_t^i < r_i$ and $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_N)$. If a SS satisfies $r_i \in \{0, 1\}$ a.s., then $\tilde{r} = r$ a.s. and we say that r is integer. The set of integer SS is denoted by $S_I(\mathcal{G}_t)$. In our case study, the solutions obtained are integer or approximately integer.

It is also convenient to have a formal definition of a Score Function.

Definition 2. *Let r be a SS, $\delta = (\delta_1, \ldots, \delta_N)^T$ a random vector and $\Theta \in \mathbb{R}$, such that $r_i = 1$ if and only if $\delta_i \leq \Theta$. Then, we say that δ is a Score Function generating r with threshold Θ.*

In order to formally pose the problem, we will assume that after manual editing, the true values of a questionnaire are obtained. Thus, we have to consider only the observed and true values. We define the *edited* statistic $X^k(r)$ as the one calculated with the values obtained after editing according to a certain choice. We can write $X^k(r) = \sum \omega_{ij}^k (x_t^{ij} + \tilde{r}_i \varepsilon_t^{ij})$.

The quality of $X^k(r)$ has to be measured according to a certain function. In this paper, we consider only the Squared Error, $(X^k(r) - X^k)^2$. This choice makes easier the theoretical analysis. It remains for future research to adapt the method for other loss functions. The value of the loss function can be written as,

$$(X^k(r) - X^k)^2 = \sum_{i,i'} \epsilon_i^k \epsilon_{i'}^k \tilde{r}_i \tilde{r}_{i'}, \tag{7}$$

where $\epsilon_i^k = \sum_j \omega_{ij}^k \varepsilon_t^{ij}$ or, in matrix form, as $(X^k(r) - X^k)^2 = \tilde{r}' E^k \tilde{r}$, with $E^k = \{E_{i,i'}^k\}_{i,i'}$ and $E_{i,i'}^k = \epsilon_i^k \epsilon_{i'}^k$. We can now state the problem of selection as an optimization problem,

$$\begin{array}{lll} [P_Q] & \max_r & \mathbf{E}[1^T \tilde{r}] \\ & \text{s.t.} & r \in S(\mathcal{G}_t), \mathbf{E}[\tilde{r}^T E^k \tilde{r}] \leq e_k^2, k = 1, \ldots, p. \end{array}$$

In section 6 we will see the solution to this problem. The vector in the cost function can be substituted for another one in case the editing work were considered different among units (e.g., if we want to reduce the burden for some respondents; this possibility is not dealt with in this paper).

Let us now analyse the expression (7). We can decompose it as,

$$(X^k(r) - X^k)^2 = \sum_i (\epsilon_i^k)^2 \tilde{r}_i + \sum_{i \neq i'} \epsilon_i^k \epsilon_{i'}^k \tilde{r}_i \tilde{r}_{i'}. \tag{8}$$

The first term in the RHS of (8) accounts for the individual impact of each error independently of its sign. In the second term the products are negative when the factors have different signs. Therefore, in order to reduce the total error, a strategy will be better if it tends to leave unedited those couples of units with different signs. The nonlinearity of the second term makes the calculations more involved. For that reason, we will also study the problem neglecting the second term.

$$\begin{array}{lll} [P_L] & \max_r & \mathbf{E}[1^T \tilde{r}] \\ & \text{s.t.} & r \in S(\mathcal{G}_t), \mathbf{E}[D^k \tilde{r}] \leq e_k^2, k = 1, \ldots, p, \end{array}$$

where $D^k = (D_1^k, \ldots, D_N^k)^T$, $D_i^k = (\epsilon_i^k)^2$

This problem is easier than P_Q because the constraints are linear. In section 5 we will see that the solution is given by a certain score function. Since there is no theoretical justification for neglecting the quadratic terms, the SS solution of the linear problem has to be empirically justified by the results obtained with real data.

5 Linear case

In this section, we analye problem $[P_L]$. The reduction to the full information case yields $f^*(r, \omega) = 1^T r$ and $g_2^*(r, \omega) = \Delta^k r$, where $\Delta^k = \mathbf{E}[D^k | \mathcal{G}_t]$. From now onwards, we write f and g_2 instead of f^* and g_2^*. Therefore, $[P_L]$ can be stated as a particular case of $[P]$ with,

$$\begin{array}{ll} x = r & \mathcal{G} = \mathcal{G}_t \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & f(r, \omega) = \mathbf{1}^T r \\ g_2(r, \omega) = (\Delta^1(\omega) r - e_1^2, \ldots, \Delta^p(\omega) r - e_p^2)^T & \end{array} \tag{9}$$

In order to avoid heavy notation, the dependence of Δ^k on ω will be implicit in the subsequent analysis. Note that in our application, f does not depend on ω.

Proposition 4. *If Δ^k has finite expectation for all k and $\mathcal{G}_t$ is countably generated, then assumptions 1–5 hold for the case defined by (9).*

Proof. For fixed r, g_2 is measurable in ω because it is a linear combination of measurable functions. Therefore, assumption 1 holds. The convexity of the

functions is due to the linearity. For the continuity condition of assumption 2 is sufficient that $\Delta^k \in L^1$, because of,

$$|\mathbf{E}[g_2(z)] - \mathbf{E}[g_2(r)]| = |\int [g_2^k(z(\omega),\omega) - g_2^k(r(\omega),\omega)]P(\mathrm{d}\omega)| \leq$$
$$\leq \int |g_2^k(z(\omega),\omega) - g_2^k(r(\omega),\omega)|P(\mathrm{d}\omega) \leq \|\Delta^k\|_{L^1}\|z-r\|_{L^\infty}$$

The conditions on f trivially hold, since it is linear and nonrandom. Assumption 3 holds with $x_0 = 0$ and assumptions 4 and 5 are obvious. □

The deterministic problem in this case can be expressed as,

$$\begin{array}{ll} \max\limits_{r} & 1^T r - \sum_k \lambda_k(\Delta^k r - e_k^2) \\ \text{s.t.} & r_i \in [0,1]. \end{array}$$

By applying the Karush–Kuhn–Tucker conditions, we get that the solution is given by,

$$r_i = \left\{ \begin{array}{lll} 1 & \text{if} & \lambda^T\Delta_i < 1 \\ 0 & \text{if} & \lambda^T\Delta_i > 1 \end{array} \right. , \tag{10}$$

where $\Delta_i = (\Delta_i^1, \ldots, \Delta_i^p)^T$. The case $\lambda^T\Delta_i = 1$ is a zero-probability event since we deal with quantitative data and then, the distribution of Δ_i is continuous. This implies,

Proposition 5. *The solution to $[P_L]$ is the SS generated by the Score Function $\delta_i = \lambda^T\Delta_i$ with threshold equal to 1.*

We describe in section 7 how to use a model for the practical computation of Δ^k. In order to estimate the dual function $\varphi(\lambda) = \mathbf{E}[L(\lambda, x_\lambda)]$ we replace expectation for the mean value over a sample as we did in section 3. However, in this case we can obtain a sample of real data instead of a simulated one because we have realizations of the variables for several periods. Thus, we can seek the optimum of $\hat{\varphi}(\lambda) = \frac{1}{h}\sum_{t=t_0}^{t_0+h-1} L_t(\lambda, r^t(\lambda))$.

Summarizing the foregoing discussion, we have proved that,

- The optimum solution to $[P_L]$ is a score-function method whose global SF is a linear combination of the local SF of the different indicators $k = 1, \ldots, p$.
- The local score function of indicator k is given by $\Delta_i^k = \mathbf{E}[(\epsilon_i^k)^2|\mathcal{G}_t]$.
- The coefficients λ^k of the linear combination are those which maximize $\varphi(\lambda)$.
- The threshold is 1.

6 Quadratic case

The outline of this section is similar to that of the previous one, but the quadratic problem poses some further difficulties, in particular, that the constraints are not convex. Therefore, we will replace them by some convex ones in such a way that under some assumptions the solutions remain the same.

Lemma 2. *The following identity holds,*

$$\mathbf{E}[\tilde{r}^T E^k \tilde{r}|\mathcal{G}_t] = r^T \Gamma^k r + (\Delta^k)^T r \tag{11}$$

where $\Gamma^k = \{\Gamma^k_{ij}\}_{ij}$ *and,*

$$\Gamma^k_{ij} \quad = \quad \begin{cases} \mathbf{E}[\epsilon^k_i \epsilon^k_j|\mathcal{G}_t] & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j \end{cases}$$

Proof. Since $(\tilde{r}_i)^2 = \tilde{r}_i$ can write,

$$\mathbf{E}[\tilde{r}^T E^k \tilde{r}|\mathcal{G}_t] = \sum_i \mathbf{E}[(\epsilon^k_i)^2 \tilde{r}_i|\mathcal{G}_t] + \sum_{i \neq i'} \mathbf{E}[\epsilon^k_i \epsilon^k_{i'} \tilde{r}_i \tilde{r}_{i'}|\mathcal{G}_t] \tag{12}$$

If we define $\mathcal{G}^*_t = \mathcal{G}_t \times \sigma(\chi^i) \times \sigma(\chi^{i'})$, by using that $\mathbf{E}[\cdot|\mathcal{G}_t] = \mathbf{E}[\mathbf{E}[\cdot|\mathcal{G}^*_t]|\mathcal{G}_t]$, we can write the right hand side of (12) as,

$$\sum_i \mathbf{E}[\mathbf{E}[(\epsilon^k_i)^2 \tilde{r}_i]|\mathcal{G}^*_t]|\mathcal{G}_t] + \sum_{i \neq i'} \mathbf{E}[\mathbf{E}[\epsilon^k_i \epsilon^k_{i'} \tilde{r}_i \tilde{r}_{i'}]|\mathcal{G}^*_t]|\mathcal{G}_t] \tag{13}$$

Since $\tilde{r}_i$ and $\tilde{r}_{i'}$ are $\mathcal{G}^*_t-$measurable, (13) can be expressed as,

$$\sum_i \mathbf{E}[\tilde{r}_i \mathbf{E}[(\epsilon^k_i)^2]|\mathcal{G}^*_t]|\mathcal{G}_t] + \sum_{i \neq i'} \mathbf{E}[\tilde{r}_i \tilde{r}_{i'} \mathbf{E}[\epsilon^k_i \epsilon^k_{i'}]|\mathcal{G}^*_t]|\mathcal{G}_t]$$

but χ^i and $\chi^{i'}$ are independent from $\mathcal{F}$, so $\mathbf{E}[\epsilon^k_i \epsilon^k_{i'}|\mathcal{G}^*_t] = \mathbf{E}[\epsilon^k_i \epsilon^k_{i'}|\mathcal{G}_t] = \Gamma^k_{ii'}$ and $\mathbf{E}[(\epsilon^k_i)^2|\mathcal{G}^*_t] = \mathbf{E}[(\epsilon^k_i)^2|\mathcal{G}_t] = \Delta^k_i$. Now, $\Gamma^k_{ii'}$ and Δ^k_i are $\mathcal{G}_t-$measurable. Finally, using that $E[\tilde{r}_i|\mathcal{G}] = r_i$ and $E[\tilde{r}_i \tilde{r}_{i'}|\mathcal{G}] = r_i r_{i'}$ we get (11). □

Therefore, $[P_Q]$ is a particular case of $[P]$ with,

$$\begin{array}{ll} x = r & \mathcal{G} = \mathcal{G}_t \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & f(r) = \mathbf{1}^T r \\ g_2(r,\omega) = (r^T \Gamma^1 r + (\Delta^1)^T r - e^2_1, \ldots, r^T \Gamma^p r + (\Delta^p)^T r - e^2_p)^T & \end{array} . \tag{14}$$

Where the dependence of the conditional moments on ω is again implicit. Unfortunately, the matrices Γ^k are indefinite and thus the constraints are not convex. We will overcome this difficulty by using the following lemma.

Lemma 3. *Let* $\bar{g}_2$ *be a function such that* $\forall r \in S_I(\mathcal{G}), \bar{g}_2(r) = g_2(r)$ *and* $\forall r \in S(\mathcal{G}), \bar{g}_2(r) \leq g_2(r)$*, and let* $[P'_Q]$ *be the problem obtained from* $[P_Q]$ *substituting* g_2 *for* $\bar{g}_2$*. Then, if* r *is a solution to* $[P'_Q]$ *and* $r \in S_I(\mathcal{G})$*, then* r *is a solution to* $[P_Q]$*.*

Proof. Let us assume that r is a solution to $[P'_Q]$ and it is integer. We know that r satisfies $\mathbf{E}[\bar{g}_2(r)] \leq e_k^2$ for $k = 1, \ldots, p$. Then, $\mathbf{E}[g_2(r)] \leq e_k^2$, so r is satisfies the constraints of $[P_Q]$. Let $s \in S(\mathcal{G}_t)$ such that $\mathbf{E}[g_2(r)] \leq e_k^2$. Since $\mathbf{E}[\bar{g}_2(r)] \leq \mathbf{E}[g_2(r)]$ then $\mathbf{E}[\bar{g}_2(r)] \leq e_k^2$ and then, $\mathbf{E}[\mathbf{1}^T s] \leq \mathbf{E}[\mathbf{1}^T r]$. □

We may consider for example the two following possibilities,

(i) $\bar{g}_2(r) = r^T \Sigma^k r$, where $\Sigma_{ij}^k = \mathbf{E}[\epsilon_i^k \epsilon_j^k | \mathcal{G}_t]$

(ii) $\bar{g}_2(r) = r^T M^k r + (v^k)^T r$, where $M_{ij}^k = m_i^k m_j^k$, $m_i^k = \mathbf{E}[\epsilon_i^k | \mathcal{G}_t]$, $v_i^k = \mathbf{V}[\epsilon_i^k | \mathcal{G}_t]$.

The case (ii), can be used only under the assumption that $\mathbf{E}[\epsilon_i^k \epsilon_j^k | \mathcal{G}_t] = m_i^k m_j^k$ for $i \neq j$ and this will be the one used in our application (section 8). Lemma 3 has practical relevance if we check that the solutions of $[P'_Q]$ are integer. We will show that this approximately holds in our application.

Problem $[P'_Q]$ is a case of $[P]$ with,

$$\begin{array}{ll} x = r & f(r) = \mathbf{1}^T r \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & \\ g_2(r, \omega) = (r^T A^1 r + (b^1)^T r - e_1^2, \ldots, r^T A^p r + (b^p)^T r - e_p^2)^T & \end{array} \tag{15}$$

Where $A^k = \Sigma^k, b^k = 0$ or $A^k = M^k, b^k = v^k$. Since A^k are positive semidefinite, we can state,

Proposition 6. *If A^k and b^k have finite expectation for all k and $\mathcal{G}_t$ is countably generated, then assumptions 1–5 hold for the case defined by (15).*

Proof. The arguments of proposition 4 can be easily adapted to the quadratic case given that the matrices that appear in the definition of g_2 are positive semidefinite. For the continuity condition of assumption 2 note that,

$$\begin{aligned} |g_2^k(z(\omega), \omega) - g_2^k(r(\omega), \omega)| &= |z^T A^k z + (b^k)^T z - r^T A^k r - (b^k)^T r| \leq \\ &\leq |z^T A^k (z - r)| + |r^T A^k (z - r)| + |(b^k)^T (z - r)| \end{aligned}$$

Then,

$$|\mathbf{E}[g_2(z)] - \mathbf{E}[g_2(r)]| \leq \left\{ \left[\|z\|_{L^\infty} + \|r\|_{L^\infty} \right] \|A^k\|_{L^1} + \|b^k\|_{L^1} \right\} \|z - r\|_{L^\infty}$$

If $z \to r$ in L^∞, the right hand side of the inequality above converges to zero. As in proposition 4, the conditions on f trivially hold. □

As in the linear case, $r(\lambda)$ is obtained solving a deterministic optimization problem, in this case a quadratic programming problem.

$$\max_r \quad 1^T r - \textstyle\sum_k \lambda_k (\bar{g}_2^k(r) - e_k^2) \tag{16}$$

$$\text{s.t.} \qquad r_i \in [0, 1]. \tag{17}$$

An important difference with respect to the linear case is that the problem above does not explicitly provide a Score Function generating the SS as when applying the Karush–Kuhn–Tucker conditions in section 5.

We describe in section 7 a practical method to obtain M^k, Σ^k and v^k. It is easy to solve $[P_D(\lambda, \omega)]$ in the linear case, but for large sizes (in our case $N > 10,000$), the quadratic programming problem becomes computationally heavy if solved by traditional methods. For $\bar{g}_2$ defined as in (ii), we can take advantage of the low rank of the matrix in the objective function to propose (appendix A) an approximate method to solve it efficiently. In our real data study, we have checked the performance of this method and the results are presented in subsection 8.1.

7 Model-based conditional moments

The practical application of the results in previous sections requires a method to compute the conditional moments of the error with respect to $\mathcal{G}_t$. In this section, we drop the index j to reduce the complexity of the notation, but the results can be adapted to the case of several variables per questionnaire.

Let $\mathcal{H}_t$ be a $\sigma-$field generated by all the information available at time t with the exception of $\tilde{x}_t^i$. Then, $\mathcal{G}_t = \sigma(\tilde{x}_t^i, \mathcal{H}_t)$. Let $\hat{x}_t^i = \tilde{\pi}(x_t^i)$ be a predictor computed using the information in $\mathcal{H}_t$, that is a $\mathcal{H}_t$-measurable random variable optimal in some way decided by the analyst. The prediction error is denoted by $\xi_t^i = x_t^i - \hat{x}_t^i$

We assume that,

Assumption 6. *ξ_t^i and η_t^i are distributed as a bivariate Gaussian with zero mean, variances ν_i^2 and σ_i^2 and correlation γ_i.*

Assumption 7. *$\varepsilon_t^i = \eta_t^i e_t^i$, where e_t^i is a Bernoulli variable that equals 1 or 0 with probabilities p and $1-p$ and it is independent of ξ_t^i and η_t^i.*

Assumption 8. *ξ_t^i, η_t^i and e_t^i are jointly independent of $\mathcal{H}_t$.*

With these assumptions, the conditional moments of the error with respect to $\mathcal{G}_t$ are functions of the sole variable $u_t^i = \hat{x}_t^i - \tilde{x}_t^i$, that is, the difference between the predicted and the observed values. In the next proposition we will also drop i and t in order to simplify notation.

Proposition 7. *Under the assumptions 6–8, it holds,*

$$E[\varepsilon|\mathcal{G}] = \frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} u\zeta \tag{18}$$

$$E[\varepsilon^2|\mathcal{G}] = \left[\frac{\sigma^2\nu^2(1-\gamma^2)}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} + \left(\frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu}\right)^2 u^2\right]\zeta, \tag{19}$$

where,

$$\zeta = \frac{1}{1 + \frac{1-p}{p}\left(\frac{\nu^2}{\sigma^2+\nu^2+2\gamma\sigma\nu}\right)^{-1/2} \exp\{-\frac{u^2(\sigma^2+2\gamma\sigma\nu)}{2\nu^2(\sigma^2+\nu^2+2\gamma\sigma\nu)}\}}. \tag{20}$$

Proof. First of all, $E[\varepsilon^\alpha|u] = E[e\eta^\alpha|u] = E[E[e\eta^\alpha|e,u]|u]$. Since e is $\sigma(e,u)-$measurable, $E[e\eta^\alpha|e,u] = E[\eta^\alpha|e,u]e$. We can prove that,

$$E[\eta^\alpha|e,u] = \begin{cases} \psi(u) & e=1 \\ \sigma^2 & e=0 \end{cases},$$

where ψ is such that $\mathbf{E}[\eta^\alpha|\xi+\eta] = \psi(\xi+\eta)$. Therefore, $E[\eta^\alpha|e,u]e = \psi(u)e$, and then, $E[E[e\eta^\alpha|e,u]|u] = \psi(u)E[e|u]$.

It remains to compute $E[e|u]$ and $\psi(u)$ for $\alpha = 1, 2$. For this purpose, we can use the properties of the Gaussian distribution. If (x,y) is a Gaussian-distributed random vector with zero mean and covariance matrix $(\sigma_{ij})_{i,j\in\{x,y\}}$. Then, the conditional distribution $f(y|x)$ is a Gaussian with mean and variance,

$$E[y|x] = \frac{\sigma_{xy}}{\sigma_{xx}}x \qquad V[y|x] = \sigma_{yy} - \frac{{\sigma_{xy}}^2}{\sigma_{xx}}.$$

Then,

$$E[y^2|x] = \sigma_{yy} + \frac{{\sigma_{xy}}^2}{\sigma_{xx}}\left(\frac{x^2}{\sigma_{xx}} - 1\right)$$

Now, we can apply the relations above to $y = \eta$ and $x = u = \eta + \xi$. Then, $\sigma_{xx} = \sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $\sigma_{yy} = \sigma^2$, $\sigma_{xy} = \sigma^2 + \gamma\sigma\nu$. Thus, for $\alpha = 1$,

$$\psi(u) = \frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu}u,$$

and for $\alpha = 2$,

$$\psi(u) = \sigma^2 + \frac{(\sigma^2 + \gamma\sigma\nu)^2}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu}\left(\frac{u^2}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} - 1\right).$$

Let us now compute $\zeta = E[e|u] = P[e=1|u]$. By an argument similar to Bayes's theorem it can be proved that ζ is equal to $P[e=1]f(u|e=1)/f(u)$, where $f(u|e=1)$ is a zero-mean Gaussian density function with variance $\sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $f(u)$ is a mixture of two Gaussians with variances $s_1^2 = \sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $s_2^2 = \nu^2$ and probabilities p and $1-p$ respectively. Hence,

$$\zeta = p\frac{(2\pi s_1^2)^{-1/2}\exp\{-\frac{u^2}{2s_1^2}\}}{p(2\pi s_1^2)^{-1/2}\exp\{-\frac{u^2}{2s_1^2}\} + (1-p)(2\pi s_2^2)^{-1/2}\exp\{-\frac{u^2}{2s_2}\}}.$$

After simplifying it yields,

$$\zeta = \frac{1}{1 + \frac{1-p}{p}\left(\frac{\nu^2}{\sigma^2+\nu^2+2\gamma\sigma\nu}\right)^{-1/2}\exp\{-\frac{u^2(\sigma^2+2\gamma\sigma\nu)}{2\nu^2(\sigma^2+\nu^2+2\gamma\sigma\nu)}\}}.$$

Finally, since ξ, η and e are independent of $\mathcal{H}_t$, we can conclude by noting that $E[\varepsilon^2|u] = E[\varepsilon^2|u,\mathcal{H}_t] = E[\varepsilon^2|\mathcal{G}_t]$. □

8 Case Study

In this section, we present the results of the application of the methods described in this paper to the data of the Turnover/New Orders Survey. Monthly data from about $N = 13,500$ units are collected. In the moment of the study, data from January 2002 to September 2006 were available (t ranges from 1 to 57). Only two of the variables requested in the questionnaires are considered in our study, namely, Total Turnover and Total New Orders ($q = 2$). The total Turnover of unit j at period t is x_t^{i1} and Total New Orders is x_t^{i2}. These two variables are aggregated separately to obtain the two indicators, so $p = 2$ and $\omega_{i2}^1 = \omega_{i1}^2 = 0$.

We need a model for the data in order to apply proposition 7 and obtain the conditional moments. Since the variables are distributed in a strongly asymmetric way, we use their logarithm transform, $y_t^{ij} = \log(x_t^{ij} + m)$, where m is a positive constant adjusted by maximum likelihood ($m \approx 10^5$€). The conditional moments of the original variable can be recovered exactly by using the properties of the log-normal distribution or approximately by using a first-order Taylor expansion, yielding $\mathbf{E}[(\tilde{x}_t^{ij} - x_t^{ij})^2|\mathcal{G}_t] \approx (\tilde{x}_t^{ij} - m)^2\mathbf{E}[(\tilde{y}_t^{ij} - y_t^{ij})^2|\mathcal{G}_t]$. In our study, we used the approximate version. We found that if $\tilde{x}_t^{ij} - m$ is replaced by an average of the last 12 values of $\tilde{x}_t^{ij}$, the estimate becomes more robust against very small values of $\tilde{x}_t^{ij} - m$.

The model applied to the transformed variables is very simple. We assume that the variables x_t^{ij} are independent across (i, j) and for any pair (i, j), we choose among the following simple models.

$$(1 - B)y_t^{ij} = a_t \tag{21}$$

$$(1 - B^{12})y_t^{ij} = a_t \tag{22}$$

$$(1 - B^{12})(1 - B)y_t^{ij} = a_t \tag{23}$$

where B is the backshift operator $Bu_t = u_{t-1}$ and a_t are white noise processes. We obtain the residuals $\hat{a}_t$ and then select the model which produces lesser mean of squared residuals, $\sum \hat{a}_t^2/(T - r)$, where r is the maximum lag in the model. With this model, we compute the prediction $\hat{y}_t^{ij}$ and the prediction standard deviation ν_{ij}. The *a priori* standard deviation of the observation errors and the error probability are considered constant across units (that is possible because of the logarithm transformation). We denote them by σ_j and p_j with $j = 1, 2$ and they are estimated using historical data of the survey.

A database is maintained with the original collected data and subsequent versions after eventual corrections due to the editing work. Thus, we consider the first version of the data as *observed* and the last one as *true*. The coefficient γ_i is assumed zero. Once we have computed σ_j, p_j, ν_{ij} and u_t^{ij}, proposition 7 can be used to obtain the conditional moments and then, Δ^k, Σ^k and v^k.

λ	mean $1^T r$	mean $1^T \lvert r - r_{app} \rvert$	λ	mean $1^T r$	mean $1^T \lvert r - r_{app} \rvert$
10^2	592.8	0.16	10^9	235.8	0.25
10^4	587.6	1.12	10^{10}	108.7	0.08
10^6	555.3	1.13	10^{11}	44.0	0.07
10^8	386.3	0.51	10^{12}	23.8	0.08

Table 2: Comparison between the exact (r) and approximate (r_{app}) quadratic methods.

8.1 Accuracy of the Approximate Method to the Quadratic Problem.

Before assessing the efficiency of the selection, we have used the data to check that the approximate method to solve the quadratic problem does not produce a significant disturbance of the solutions. We have compared the approximate solutions to the ones obtained by a usual quadratic programming approach. For this purpose, we have used the function *quadprog* of the mathematical pack MATLAB. For the whole sample, *quadprog* does not converge in a reasonable time —that is the reason why the approximate method is required— so we have extracted for comparison a random subsample of 5% (roughly over 600 units) and we have solved the problem $[P_Q]$ for a range of values of λ with the exact and approximate methods. In table 2, we present for the different values of λ, the average of the number of units edited using the exact method and a measure of the difference between the two methods. We also have used this data to check the validity of the assumption that the solutions (of the exact method) are integer. For this purpose, we computed $\sum_i \min\{r_i, 1 - r_i\}$, whose value never exceeded 1, while the number of units for which $\min\{r_i, 1 - r_i\} > 10^{-3}$ was at most 2. Therefore, the solution can be considered as approximately integer.

8.2 Expectation Constraints

We will now check that the expectation constraints in $[P_L]$ and $[P_Q]$ are effectively satisfied. In order to do this, for $l = 1, \ldots, b$ with $b = 20$ we solve the optimization problem with the variance bounds $e_{1l}^2 = e_{2l}^2 = e_l^2 = [s_0^{((l-1)/(b-1))} s_1^{((b-l)/(b-1))}]^2$. The range of standard deviations goes from $s_0 = 0.025$ to $s_1 = 1$.

The expectation of the dual function is estimated using a $h-$length batch of real data. For any period from October 2005 to September 2006 and for any $l = 1, \ldots, b$, a selection $r(t, l)$ is obtained according the bound e_k^2. The average across t of the remaining squared errors is thus computed as,

$$\hat{e}_{kl}^2 = \frac{1}{12} \sum_{t=t_0}^{t_0+11} r(t,l)^T E^k r(t,l)$$

We repeated these calculations for $h = 1, 3, 6$ and 12 both using the linear and the quadratic versions. The results are arranged in tables 4 to 11. For

	Turnover		Orders	
	E_1	E_2	E_1	E_2
δ^0	0.43	0.44	1.16	1.33
δ^1	0.30	0.38	0.36	0.45
δ^2	0.21	0.26	0.28	0.37

Table 3: Comparison of score functions.

each l we present the average number of units edited, the desired bound and for $k = 1, 2$, the quotient $\hat{e}_{kl}/e_{kl}$. In every case, there is a tendency to underestimate the error when the constraints are smaller and to overestimate it when the bounds are larger. The quadratic method produces better results with respect to the bounds but at the price of editing more units.

8.3 Comparison of Score Functions

We intend to compare the performance of our method to that of the score-function described in [6], $\delta_i^0 = \omega_i|\tilde{x}^i - \hat{x}^i|$, where $\hat{x}_i$ is a prediction of x according to a certain criterion. The author proposes to use the last value of the same variable in previous periods. We have also considered the score function δ^1 defined as δ^0 but using the forecasts obtained through the models in (21)–(23). Finally, δ^2 is the score function computed using (21)–(23) and proposition 7. The global SF is just the sum of the two local ones. We will measure the effectiveness of the score functions by $E_l^j = \sum_n E_l^j(n)$, with,

$$E_1^j(n) = \sum_{i \geq n}^{N} (\omega_i^j)^2 (\tilde{x}^{ij} - x^{ij})^2 \qquad E_2^j(n) = \Big[\sum_{i \geq n}^{N} \omega_i^j (\tilde{x}^{ij} - x^{ij}) \Big]^2,$$

where we consider units arranged in descending order according to the corresponding score function. These measures can be interpreted as estimates of the remaining error after editing the n first units. The difference is that $E_1^j(n)$ is the aggregate squared error and $E_2^j(n)$ is the squared aggregate error. Thus, $E_2^j(n)$ is the one that has practical relevance, but we also include the values of $E_2^j(n)$ because in the linear problem $[P_L]$, it is the aggregate squared error which appears in the left side of the expectation constraints. In principle, it could happen that our score function was optimal for the $E_1^j(n)$ but not for $E_2^j(n)$. Nevertheless, the results in table 3 show that δ^2 is better measured both ways.

9 Conclusions.

We have described a theoretical framework to deal with the problem of selective editing, defining the concept of selection strategy. We describe the search for

an adequate selection strategy as an optimization problem. This problem is a linear optimization problem with quadratic constraints. We show that the score function approach is the solution to the problem with linear constraints. We also show how to solve the quadratic problem.

The score function obtained outperforms a reference SF. Both the linear and the quadratic versions of our method produce selection strategies that satisfy approximately the constraints but for small values of the constraint. The quadratic method seems to be more conservative and then, the bounds are better fulfilled, but more units are edited. On the other hand, the implementation of the linear method is easier and computationally less demanding. This suggests that the quadratic method is more adequate for cases in which the bounds are critical and the linear one for cases in which timeliness is critical.

Acknowledgements

The authors wish thank José Luis Fernández Serrano and Emilio Cerdá for their help and advice.

References

[1] J. Berthelot and M. Latouche. Improving the efficiency of data collection: A generic respondent follow-up strategy for economic surveys. *J. Bus. Econom. Statist.*, 11:417–424, 1993.

[2] D. Bertsimas and M. Sim. The price of rebustness. *Oper. Res.*, 52:35–53, 2004.

[3] P. Billingsley. Probability and measure. John Wiley and Sons, New York, 1995.

[4] A. M. Croicu and M. Y. Hussaini. On the expected optimal value and the optimal expected value. *Appl. Math. Comput.*, 180:330–341, 2006.

[5] L. Granquist. The new view on editing. *International Statistics Review*, 65:381–387, 1997.

[6] D. Hedlin. Score functions to reduce business survey editing at the u.k. office for national statistics. *Journal of Official Statistics*, 19:177–199, 2003.

[7] M. Latouche and J. Berthelot. Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8:389–400, 1992.

[8] D. Lawrence and R. McKenzie. The general application of significance editing. *Journal of Official Statistics*, 16:243–253, 2000.

[9] D. G. Luenberger. Optimization by vector space methods. John Wiley and Sons, New York, 1969.

[10] S. M. Robinson. Analysis of sample-path optimization. *Math. Oper. Res.*, 21:513–528, 1996.

[11] R. T. Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture notes in Math*, pages 157–207. Springer, Berlin, 1976.

[12] R. T. Rockafellar and R. J. B. Wets. Variational analysis. Springer, New York, 1998.

[13] W. Rudin. Real and complex analysis. McGraw-Hill, New York, 3rd edition, 1987.

[14] A. Shapiro and T. Homem-de-Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastics programs. *SIAM J. Optim.*, 11:70–86, 2000.

Appendix

A Approximate method for the quadratic problem

The Karush–Kuhn–Tucker conditions applied to the problem (16)–(17) with $\bar{g}_2^k(r) = r^T M^k r + (v^k)^T r - e_k^2$ imply that in the optimum, the following holds,

$$2\sum_k \lambda_k m_i^k (m^k)^T r + v_i^k - 1 = \mu_i^+ - \mu_i^- \quad \mu_i^+, \mu_i^- \geq 0$$
$$\mu_i^+(1 - r_i) = 0 \quad \mu_i^- r_i = 0$$

where $m^k = (m_1^k, \ldots, m_N^k)^T$. The relations above hold when,

$$\begin{array}{l} 2\sum_k \lambda_k m_i^k (m^k)^T r + v_i^k - 1 > 0 \quad \text{if} \quad r_i = 1 \\ 2\sum_k \lambda_k m_i^k (m^k)^T r + v_i^k - 1 < 0 \quad \text{if} \quad r_i = 0 \end{array}$$

Let us assume that we know $\alpha = [(m^1)^T r, \ldots, (m^p)^T r]^T = Mr$, where $M = (m^1, \ldots, m^p)^T$. Then, we can built $r(\alpha)$ as,

$$r_i = \begin{cases} 1 & \text{if} \quad 2\sum_k \lambda_k m_i^k \alpha_k + v_i^k > 1 \\ 0 & \text{if} \quad 2\sum_k \lambda_k m_i^k \alpha_k + v_i^k < 1 \end{cases}$$

If $\alpha = Mr(\alpha)$, then $r(\alpha)$ is a solution. We can solve approximately the fixed-point problem by minimizing $\|\alpha - Mr(\alpha)\|^2$. In our applications p is typically small, so the dimension of the optimization problem has been strongly reduced.

B Tables

Table 4: Error bounds of the linear version (h=1).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.04	3.15	397.0	0.1742	0.85	1.59	29.9
0.0304	1.87	2.72	312.2	0.2116	0.75	1.29	21.1
0.0369	1.61	2.28	245.8	0.2569	0.65	1.05	14.1
0.0448	1.46	1.99	194.8	0.3120	0.61	0.86	9.5
0.0544	1.08	1.54	153.6	0.3788	0.56	0.73	6.1
0.0660	0.90	1.38	119.6	0.4600	0.56	0.69	3.8
0.0801	0.79	1.19	92.5	0.5585	0.61	0.60	2.1
0.0973	0.72	1.02	71.9	0.6782	0.47	0.51	1.3
0.1182	0.74	1.22	54.5	0.8235	0.39	0.42	0.8
0.1435	0.88	1.63	40.9	1.0000	0.32	0.34	0.8

Table 5: Error bounds of the linear version (h=3).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.43	3.74	427.9	0.1742	1.26	1.29	33.7
0.0304	2.45	3.35	338.4	0.2116	1.18	1.18	24.3
0.0369	2.49	3.09	266.7	0.2569	1.00	1.00	17.5
0.0448	2.36	2.39	208.8	0.3120	0.87	0.81	12.0
0.0544	1.86	2.39	165.2	0.3788	0.88	0.70	8.1
0.0660	1.66	1.99	132.6	0.4600	0.79	0.59	5.4
0.0801	1.60	1.68	102.2	0.5585	0.67	0.51	3.1
0.0973	1.50	1.46	79.0	0.6782	0.56	0.44	1.8
0.1182	1.50	1.38	60.9	0.8235	0.43	0.36	1.1
0.1435	1.41	1.24	46.0	1.0000	0.35	0.31	0.7

Table 6: Error bounds of the linear version (h=6).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.89	3.20	414.8	0.1742	0.97	1.31	30.5
0.0304	1.88	2.63	327.8	0.2116	0.82	1.29	20.7
0.0369	2.04	2.43	257.6	0.2569	0.71	1.09	15.0
0.0448	1.83	2.05	202.0	0.3120	0.80	0.85	10.3
0.0544	1.58	1.87	157.7	0.3788	0.74	0.77	6.8
0.0660	1.11	1.59	125.7	0.4600	0.71	0.72	4.7
0.0801	1.10	1.25	95.9	0.5585	0.57	0.57	2.7
0.0973	0.96	1.01	73.2	0.6782	0.47	0.46	1.7
0.1182	0.96	1.18	56.2	0.8235	0.42	0.38	1.2
0.1435	0.96	1.12	41.8	1.0000	0.30	0.31	0.5

Table 7: Error bounds of the linear version (h=12).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.56	3.38	430.4	0.1742	0.76	0.86	31.3
0.0304	1.08	2.63	340.7	0.2116	0.64	1.09	20.0
0.0369	1.02	1.77	268.4	0.2569	0.97	0.91	14.0
0.0448	1.00	1.30	207.9	0.3120	0.79	0.65	7.6
0.0544	0.79	1.26	161.0	0.3788	0.78	0.64	4.4
0.0660	0.72	0.98	129.4	0.4600	0.78	0.65	2.7
0.0801	0.84	0.78	96.4	0.5585	0.64	0.48	1.3
0.0973	1.15	0.60	76.1	0.6782	0.53	0.40	0.9
0.1182	0.96	0.64	57.4	0.8235	0.43	0.33	0.7
0.1435	0.83	0.68	41.1	1.0000	0.36	0.27	0.6

Table 8: Error bounds of the quadratic version (h=1).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.41	4.12	507.6	0.1742	0.99	0.82	265.0
0.0304	1.40	3.01	655.8	0.2116	1.95	0.86	134.8
0.0369	1.60	2.57	540.8	0.2569	0.98	0.86	54.3
0.0448	1.32	2.12	541.3	0.3120	1.35	0.73	36.5
0.0544	0.80	1.71	593.5	0.3788	0.68	0.63	29.2
0.0660	1.06	1.67	469.8	0.4600	0.55	0.52	20.8
0.0801	0.81	1.44	460.6	0.5585	0.46	0.73	19.8
0.0973	0.98	1.21	275.7	0.6782	0.64	0.59	5.9
0.1182	1.19	1.05	150.8	0.8235	0.53	0.52	5.4
0.1435	1.21	0.99	272.6	1.0000	0.44	0.41	1.9

Table 9: Error bounds of the quadratic version (h=3).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.58	2.41	650.3	0.1742	1.24	0.90	85.4
0.0304	1.44	1.63	563.1	0.2116	1.14	0.86	42.3
0.0369	1.27	1.81	589.9	0.2569	0.94	0.73	38.1
0.0448	0.96	1.30	507.3	0.3120	0.80	0.85	26.3
0.0544	0.98	1.66	388.5	0.3788	0.61	0.70	20.6
0.0660	1.45	1.49	298.8	0.4600	0.53	0.49	23.9
0.0801	1.65	1.61	231.4	0.5585	0.59	0.60	15.1
0.0973	1.35	0.82	171.2	0.6782	0.47	0.46	4.3
0.1182	1.03	0.62	152.4	0.8235	0.31	0.37	3.6
0.1435	1.25	0.91	101.5	1.0000	0.31	0.32	2.7

Table 10: Error bounds of the quadratic version (h=6).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.86	2.50	578.4	0.1742	1.24	0.92	77.8
0.0304	1.46	2.02	605.9	0.2116	0.99	0.76	59.8
0.0369	1.42	1.82	401.9	0.2569	0.92	0.67	35.4
0.0448	0.95	1.35	477.3	0.3120	0.71	0.56	27.7
0.0544	0.96	1.17	390.5	0.3788	0.64	0.66	28.0
0.0660	1.41	1.28	278.8	0.4600	0.71	0.59	13.3
0.0801	1.31	1.30	283.4	0.5585	0.56	0.57	6.8
0.0973	0.99	0.84	225.8	0.6782	0.46	0.47	6.8
0.1182	1.32	1.22	140.3	0.8235	0.39	0.38	2.9
0.1435	1.30	0.95	93.1	1.0000	0.32	0.46	1.6

Table 11: Error bounds of the quadratic version (h=12).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.38	1.91	703.4	0.1742	1.19	0.95	83.3
0.0304	1.13	1.77	620.5	0.2116	0.99	0.71	69.3
0.0369	0.93	1.78	595.8	0.2569	0.82	0.57	50.8
0.0448	0.83	1.54	515.7	0.3120	0.73	0.51	46.8
0.0544	0.66	1.38	456.6	0.3788	0.59	0.45	27.8
0.0660	0.84	1.27	422.5	0.4600	0.55	0.51	18.0
0.0801	1.09	1.40	284.6	0.5585	0.58	0.61	10.8
0.0973	1.21	1.05	223.3	0.6782	0.48	0.48	4.5
0.1182	1.51	1.16	120.3	0.8235	0.39	0.40	1.3
0.1435	1.19	0.81	86.1	1.0000	0.32	0.32	1.2