CONSTRAINED ESTIMATION AND THE THEOREM OF KUHN-TUCKER

ORI DAVIDOV

Received 11 July 2004; Accepted 11 January 2005

We explore several important, and well-known, statistical models in which the estimation procedure leads naturally to a constrained optimization problem which is readily solved using the theorem of Kuhn-Tucker.

1. Introduction and motivation

There are many statistical problems in which the parameter of interest is restricted to a subset of the parameter space. The constraint(s) may reflect prior knowledge about the value of the parameter, or, may be a device used to improve the statistical properties of the estimator. Estimation and inferential procedures for such models may be derived using the theorem of Kuhn-Tucker (KT). The theorem of KT is a theorem in nonlinear programming which extends the method of Lagrange multipliers to inequality constraints. KT theory characterizes the solution(s) to general constrained optimization problems. Often, this characterization yields an algorithmic solution. In general, though, this is not the case and the theorem of KT is used together with other tools or algorithms. For example, if the constraints are linear or convex, then the tools of convex optimization (Boyd and Vandenberghe [2]) may be used; of these linear and quadratic programming are best known. More generally, interior point methods, a class of iterative methods in which all iterations are guaranteed to stay within the feasible set, may be used. Within this class, Lange [12] describes the adaptive barrier method with statistical applications. Geyer and Thompson [6] develop a Monte-Carlo method for constrained estimation based on a simulation of the likelihood function. Robert and Hwang [16] develop the prior feedback method. They show that the constrained estimator may be viewed as the limit of a sequence of formal Bayes estimators. The method is implemented using MCMC methodology. In some situations constrained problems may be reduced to isotonic regression problems. A variety of algorithms for solving isotonic regression are

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 92970, Pages 1–13
DOI 10.1155/JAMDS/2006/92970

discussed by Robertson et al. [17]; PAVA, to be discussed later, and its generalizations and the *min-max* and *max-min* formulas are perhaps the best known.

In this communication it is shown that KT theory is particularly attractive when the unconstrained estimation problem is easily solved. Thus it is an ideal method for a broad class of statistical models derived from the exponential family. We introduce KT theory and apply it in three interesting and important statistical problems, namely *ridge regression, order-restricted statistical inference, and bioequivalence*. KT theory has been applied to other statistical problems. For example, Lee [13, 14] and Mortaza and Bentler [11] used KT theory to estimate covariance matrices with constrained structure. Linear models with positivity constraints have been studied by among others, Liew [15] and Wang et al.[20]. The goal of this communication is to acquaint a broad readership with KT theory and demonstrate its usefulness by providing new insights, and further developments, in the study of some well-known and practically important problems.

2. The theorem of Kuhn and Tucker

We start with the standard set up. Let $\Theta \subseteq \mathbb{R}^p$ be the parameter space and let $l(\theta)$ be the objective function we wish to maximize. In most applications $l(\theta) = \log f(x;\theta)$ is simply the log-likelihood. Often we seek the maximizer of $l(\theta)$ over a subset of Θ characterized by $m \geq 1$ inequality constraints $c_1(\theta) \geq 0, \ldots, c_m(\theta) \geq 0$. The set $\mathcal{F} = \{\theta \in \Theta \mid c_1(\theta) \geq 0, \ldots, c_m(\theta) \geq 0\}$ is called the *feasible* set. Formally, our goal is to find

$$\hat{\theta} = \arg\max_{\theta \in \mathcal{F}} l(\theta), \tag{2.1}$$

where the "argmax" notation simply indicates that $\hat{\theta}$ is the value which maximizes $l(\theta)$ on $\mathcal{F}$. The functions $l(\theta)$ and $c_i(\theta)$, which map $\mathbb{R}^p$ into $\mathbb{R}$, are assumed to be continuously differentiable. Their derivatives with respect to θ, are denoted by $\nabla l(\theta)$ and $\nabla c_i(\theta)$. We start by presenting the theorem of KT and follow up with some clarifications.

Theorem 2.1. *Let $\hat{\theta}$ denote a local maximum on the feasible set and let $\mathcal{E}$ denote the set of effective constraints at $\hat{\theta}$. If the rank of the matrix $\nabla c_{\mathcal{E}}(\hat{\theta})$ is equal to the number of effective constraints, that is, if*

$$\rho\left(\nabla c_{\mathcal{E}}(\hat{\theta})\right) = |\mathcal{E}|, \tag{2.2}$$

then there is a vector $\hat{\lambda}$ for which the relationships

$$\nabla l(\hat{\theta}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla c_i(\hat{\theta}) = 0, \tag{2.3}$$

$$\hat{\lambda}_i \geq 0, \quad \hat{\lambda}_i c_i(\hat{\theta}) = 0 \quad \text{for } i = 1, \ldots, m \tag{2.4}$$

hold.

We say that the ith constraint is effective at $\hat{\theta}$ if $c_i(\hat{\theta}) = 0$. The requirement (2.2) is called the *constraint qualification*. The left-hand side of (2.2) is the rank of the derivative

matrix evaluated at the local maxima, and $|\mathscr{E}|$ is the number of effective constraints at $\hat{\theta}$. Hence (2.2) means that the derivative matrix is of full rank at the local maxima. Recall that the constraints require that $c_i(\hat{\theta}) \geq 0$. Hence (2.4) implies that if $\hat{\lambda}_i > 0$, then $c_i(\hat{\theta}) = 0$ and if $\hat{\lambda}_i = 0$, then $c_i(\hat{\theta}) > 0$. Consequently the condition (2.4) is known as *complementary slackness*. That is, if one inequality is "slack" (not strict), the other cannot be. The vector λ is known as the KT *multipliers*. The function

$$L(\theta,\lambda) = l(\theta) + \sum_{i=1}^{m} \lambda_i c_i(\theta) \tag{2.5}$$

is called the Lagrangian. In practice, local maxima are found by solving a system of equalities (2.3) and inequalities (2.4) on the feasible set, that is,

$$\begin{gathered} \nabla L(\theta,\lambda) = 0, \\ c_i(\theta) \geq 0, \quad \lambda_i \geq 0, \quad \lambda_i c_i(\theta) = 0 \quad \text{for } i = 1,\ldots,m. \end{gathered} \tag{2.6}$$

Here ∇L denotes the derivative with respect to θ. Note that the theorem of KT only gives necessary conditions for local maxima. In general, these conditions are not sufficient. However, in many statistical applications, including our examples, KT finds the unique maximizer. For a more thorough and rigorous discussion, see Sundaram [19].

3. Applications

Three applications are discussed in detail. Section 3.1 develops the ridge estimator for linear models. Our perspective on ridge regression is a bit different from the usual approach encountered throughout the statistical literature. Note that the constraints in ridge regression are usually not part of the model but a statistical device used to improve the mean squared error of the estimator. Section 3.2 deals with order-restricted inference for binary data. In this situation the values of the parameters are a priori and naturally ordered. Constrained estimation is an obvious aspect of the model. Using KT theory we develop a simple estimating procedure. We indicate how to generalize our result to the estimation of stochastically ordered distribution functions for arbitrary random variables. Finally, in Section 3.3 we develop an estimation procedure for the multitreatment bioequivalence problem. Our estimation procedure, based on KT theory, generalizes the current practice by which equivalence is assessed for two treatments at a time.

3.1. Ridge regression. Ridge regression is a well-known statistical method originally designed to numerically stabilize the estimator of the regression coefficient in the presence of multicollinearity (Hoerl and Kennard [9]). More broadly ridge regression may be viewed as a statistical shrinkage method (Gruber [7]) with multiple uses, one of which is variable selection (Hastie et al. [8]). Consider the standard linear model

$$Y = X\theta + \varepsilon, \tag{3.1}$$

where $Y^T = (y_1,\ldots,y_n)$ is the vector of outcomes, $X = ((x_{ij}))$ is the model matrix, and

$\theta^T = (\theta_1, \ldots, \theta_p)$ is the unknown parameter vector. The ridge estimator is defined by

$$\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right\} \tag{3.2}$$

for some fixed $\lambda \geq 0$. Thus the ridge estimator is a penalized least square estimator with penalty proportional to its length. Note that the ridge estimator is not equivariant under scaling. Therefore it is common to standardize the data before fitting the model; most commonly the dependent variable is centered about its sample average and the independent variables are both centered and scaled. Consequently the intercept, θ_0, is set equal to $\overline{y}$ and plays no role in (3.2). A straightforward calculation reveals that the ridge estimator is given by

$$(X^T X + \lambda I)^{-1} X^T Y. \tag{3.3}$$

Typically (3.2) is fit for a range of λ values (also known as the complexity parameter) and an "optimal" value of λ, one which reduces the empirical mean squared error, is then chosen.

Alternatively consider the following constrained estimation problem. Let

$$l(\theta) = -(Y - X\theta)^T (Y - X\theta), \qquad c(\theta) = K^2 - \theta^T \theta, \tag{3.4}$$

and find

$$\max \{ l(\theta) \mid c(\theta) \geq 0 \}. \tag{3.5}$$

In other words, find the estimator which minimizes the sum of squares over θ values within a distance of K from the origin. Clearly we may solve this optimization problem using the theorem of KT. The Lagrangian is

$$L(\theta, \lambda) = -(Y - X\theta)^T (Y - X\theta) + \lambda (K^2 - \theta^T \theta). \tag{3.6}$$

Critical points are found by solving (2.3) and (2.4) on the feasible set. It is straightforward to see that (2.3) reduces to

$$X^T Y - X^T X\theta + \lambda\theta = 0. \tag{3.7}$$

Equation (2.4) and the constraint lead to three relations

$$K^2 \geq \theta^T \theta, \qquad \lambda \geq 0, \qquad \lambda (K - \theta^T \theta) = 0. \tag{3.8}$$

The system (3.7) and (3.8) may seem, at a first glance, complicated, but in fact it is very simple. We start by noting that for any fixed value of $\hat{\lambda}$ (3.7) is linear, thus

$$\hat{\theta} = \hat{\theta}(\hat{\lambda}) = (X^T X + \hat{\lambda} I)^{-1} X^T Y. \tag{3.9}$$

At this stage complementary slackness comes in handy because it can be used to deduce the value of $\hat{\lambda}$. Note that $\hat{\theta}(0)$ is the ordinary least squares estimator. Suppose

that $\hat{\theta}\,(0)^T\hat{\theta}\,(0) \le K^2$, that is the unconstrained and constrained maxima coincide. It follows from complementary slackness that we must have $\hat{\lambda} = 0$. On the other hand if $\hat{\theta}\,(0)^T\hat{\theta}\,(0) > K^2$, then by complementary slackness we must have $\hat{\lambda} > 0$ and $K^2 - \hat{\theta}\,(\hat{\lambda})^T\hat{\theta}\,(\hat{\lambda}) = 0$. Thus we obtain the following equation for $\hat{\lambda}$:

$$Y^TX(X^TX+\hat{\lambda}I)^{-1}(X^TX+\hat{\lambda}I)^{-1}X^TY = K^2. \tag{3.10}$$

It is easily verified that the left-hand side of (3.10) is a decreasing function of $\hat{\lambda}$ therefore (3.10) has a unique solution on the set $\hat{\theta}\,(0)^T\hat{\theta}\,(0) > K^2$. To summarize,

$$(\hat{\theta},\hat{\lambda}) = \begin{cases} (\hat{\theta}\,(0),0) & \text{if } Y^TX(X^TX)^{-2}X^TY \le K^2, \\ (\hat{\theta}\,(\hat{\lambda}),\hat{\lambda}) & \text{if otherwise,} \end{cases} \tag{3.11}$$

where $\hat{\lambda}$ solves (3.10). It is easy to verify that $(\hat{\theta},\hat{\lambda})$ above satisfy (2.3) and (2.4). In addition

$$\nabla c(\theta) = -2\theta \neq 0 \quad \forall\,\theta \text{ satisfying the constraint } \theta^T\theta = K^2. \tag{3.12}$$

Therefore the constraint qualification (2.2) holds and by KT $\hat{\theta}$ must be a local maxima. Moreover by the theorem of Weierstrass, which states that a continuous function on a compact set must have a maxima on it, $l\,(\theta)$ must have a global maxima on the feasible set. Since we identified only one maxima point, it must be the global maximum and therefore $\hat{\theta}$ is the constrained MLE. More generally it is known that if the objective function is concave and the feasible set is convex, then KT provides both *necessary and sufficient* conditions to identify the global maximum provided that for some $\theta \in \mathcal{F}$, $c_i(\theta) > 0$ for all i (this requirement is known as *Slater's condition*). Clearly these conditions hold in this case.

The above analysis shows that the solution to the constrained estimation problem (3.5) is the ridge estimator. In fact our derivations clarify the relationship between the dual parameters λ and K and provide further insight to the statistical properties of the ridge estimator. The relationship $K \to \lambda$ is a function whose range and image is $\mathbb{R}_+$ the nonnegative reals. Note that if $K^2 \ge \hat{\theta}\,(0)^T\hat{\theta}\,(0)$, then $\lambda = 0$, otherwise $\lambda > 0$. In statistical terms this means that if the unconstrained estimator is within distance K from the origin, there is no need to shrink it. Clearly λ increases as K decreases. Furthermore if the θ_0, the true value, satisfies $\theta_0^T\theta_0 \le K^2$, then the constrained estimator will be consistent. Otherwise it will not. The relationship $\lambda \to K$ is a *correspondence*, not a function because although positive $\lambda's$ relate to a single value of K, the value $\lambda = 0$ relates to all K in $[\hat{\theta}\,(0)^T\hat{\theta}\,(0),\infty)$.

Viewing the ridge estimator as a solution to the optimization problem (3.5) is very appealing conceptually. It clarifies the role of λ in (3.2) and explicitly relates it to the magnitude of constraint. Furthermore it suggests some interesting statistical problems, for example, the testing of $H_0 : \theta^T\theta \le K^2$ versus the alternative that $H_1 : \theta^T\theta > K^2$ (and its dual in terms of λ) and suggests an alternative approach to calculating the large sample distribution of the ridge estimator. Relating the value of the constraint K to the sample size is also of interest. These problems will be discussed elsewhere.

3.2. Order-restricted inference. There are situations in which the parameters describing a model are naturally ordered. For a comprehensive, highly mathematical, overview of the theory of order-restricted inference and its application in a variety of settings see Robertson et al. [17] and Silvapulle and Sen [18]. Briefly, their approach to estimation under order constraints is geometrical with a strong emphasis on convexity. Our derivations are more practical in their orientation. However they are easily generalized to more complicated models. To fix ideas, consider a study relating the probability of disease with an exposure such as smoking history. Suppose that three categories of exposure are defined and that it is expected that the probability of disease increases with exposure. Let n_i denote the number of individuals in each group and let X_i be the number with disease where $X_i \sim \text{Bin}(n_i, \theta_i)$. The ordering of the exposures implies that $\theta_3 \geq \theta_2 \geq \theta_1$. Therefore the log-likelihood and constraints are

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{3} x_i \log(\theta_i) + (n_i - x_i)\log(1-\theta_i), \\ c_1(\theta) &= \theta_2 - \theta_1, \\ c_2(\theta) &= \theta_3 - \theta_2. \end{aligned} \tag{3.13}$$

Clearly θ can be estimated by applying KT. The Lagrangian is

$$L(\theta, \lambda) = l(\theta) + \lambda_1 c_1(\theta) + \lambda_2 c_2(\theta), \tag{3.14}$$

and we find solutions to

$$\nabla L = \begin{pmatrix} \dfrac{x_1}{\theta_1} - \dfrac{n_1 - x_1}{1-\theta_1} - \lambda_1 \\ \dfrac{x_2}{\theta_2} - \dfrac{n_2 - x_2}{1-\theta_2} + \lambda_1 - \lambda_2 \\ \dfrac{x_3}{\theta_3} - \dfrac{n_3 - x_3}{1-\theta_3} + \lambda_2 \end{pmatrix} = 0 \tag{3.15}$$

together with

$$\begin{aligned} \frac{\partial L}{\partial \lambda_1} = \theta_2 - \theta_1 \geq 0, \quad \lambda_1 \geq 0, \quad \lambda_1 \frac{\partial L}{\partial \lambda_1} = \lambda_1(\theta_2 - \theta_1) = 0, \\ \frac{\partial L}{\partial \lambda_2} = \theta_3 - \theta_2 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_2 \frac{\partial L}{\partial \lambda_2} = \lambda_2(\theta_3 - \theta_2) = 0. \end{aligned} \tag{3.16}$$

To find the critical points of the Lagrangian we need to solve (3.15) as well as (3.16). This system is easily solved by applying the principle of complementary slackness. The general form of the solution is summarized in Table 3.1.

Clearly, the solution is determined by which constraint(s) are effective at the optimum. For example if $\widehat{\lambda}_1 = \widehat{\lambda}_2 = 0$ (case I), then (3.16) implies that $\widehat{\theta}_1 \leq \widehat{\theta}_2 \leq \widehat{\theta}_3$ and (3.15) yields $\widehat{\theta}_i = x_i/n_i$. Thus the constraints are satisfied in the unconstrained problem as well. In statistical terms the restricted and unrestricted maximum likelihood estimates (MLEs)

Table 3.1. Solutions for the constrained estimation problem involving three-ordered binomial proportions. We label $x_{ij} = x_i + x_j$ for all $1 \le i, j \le 3$. The quantities n_{ij} are similarly defined. Clearly x_{123} is the total number of events and n_{123} is the total sample size.

	Case I	Case II	Case III	Case IV
$\widehat{\theta}_1$	$\frac{x_1}{n_1}$	$\frac{x_1}{n_1}$	$\frac{x_{12}}{n_{12}}$	$\frac{x_{123}}{n_{123}}$
$\widehat{\theta}_2$	$\frac{x_2}{n_2}$	$\frac{x_{23}}{n_{23}}$	$\frac{x_{12}}{n_{12}}$	$\frac{x_{123}}{n_{123}}$
$\widehat{\theta}_3$	$\frac{x_3}{n_3}$	$\frac{x_{23}}{n_{23}}$	$\frac{x_3}{n_3}$	$\frac{x_{123}}{n_{123}}$
$\widehat{\lambda}_1$	0	0	$\frac{n_{12}}{x_{12}}\,\frac{x_1n_2 - x_2n_1}{n_{12} - x_{12}}$	$\frac{n_{123}}{x_{123}}\,\frac{x_1n_{23} - n_1x_{23}}{n_{123} - x_{123}}$
$\widehat{\lambda}_2$	0	$\frac{n_{23}}{x_{23}}\,\frac{x_2n_3 - x_3n_2}{n_{23} - x_{23}}$	0	$\frac{n_{123}}{x_{123}}\,\frac{n_3x_{12} - x_3n_{12}}{n_{123} - x_{123}}$

coincide. Similarly if $\widehat{\lambda}_1 = 0$, $\widehat{\lambda}_2 > 0$ (case II), then (3.16) imply that $\widehat{\theta}_1 \le \widehat{\theta}_2 = \widehat{\theta}_3$. Substituting back into (3.15) we find that $\widehat{\theta}_1 = x_1/n_1$ and $\widehat{\theta}_2 = \widehat{\theta}_3 = (x_2 + x_3)/(n_2 + n_3)$. A simple substituting reveals the value of $\widehat{\lambda}_2$. Cases (III) and (IV) are similarly solved. It is easily verified that these are indeed the only solutions. In addition

$$\nabla c_1(\theta) = (-1, 1, 0), \qquad \nabla c_2(\theta) = (0, -1, 1) \tag{3.17}$$

are independent of $\widehat{\theta}$ and of full rank, both separately and together, for all θ in the feasible set, thus the constraint qualification holds and the conditions of KT are satisfied. Moreover, $l(\theta)$ is concave and the feasible set is both convex and compact. Therefore the local maxima identified must be the global maxima points. Our derivations show that the KT solutions result in the famous pool adjacent violators algorithm (PAVA), which works in the following way. Let θ_i^* denote the naive MLEs. Compare θ_1^* and θ_2^*. If $\theta_1^* \le \theta_2^*$, then set $\widehat{\theta}_1 = \theta_1^*$ and continue by comparing θ_2^* and θ_3^*, and so forth. If, however, $\theta_1^* > \theta_2^*$, then reestimate θ_1^* and θ_2^* assuming that they are equal and reassign them the value $(\theta_1^* n_1 + \theta_2^* n_2)/(n_1 + n_2)$. Continue as before treating both groups as if they were one. Note that there are six possible (3!) orderings for the unconstrained MLEs. Table 3.2 relates the ordering of the naive MLEs with the constrained ones.

Rows 1 through 3 and 6 of Table 3.2 are self-explanatory. In row 4 the unconstrained MLEs satisfy $\theta_2^* < \theta_3^* < \theta_1^*$. Recall that $\theta_1 \le \theta_2$ is required. It follows that λ_1 is positive and that the estimators for θ_1 and θ_2 are equal; their initial value is set to be $(\theta_1^* n_1 + \theta_2^* n_2)/(n_1 + n_2)$. If this value is smaller than θ_3^*, then $\widehat{\theta}_1 = \widehat{\theta}_2 < \widehat{\theta}_3$, otherwise the constraint $\theta_2 \le \theta_3$ is invoked and $\widehat{\theta}_1 = \widehat{\theta}_2 = \widehat{\theta}_3$. Similar considerations apply in row 5.

It has been noted by an associate editor that the constrained estimators are dependent whereas the unconstrained ones are independent. The degree of dependence is a function of the true parameter values. If the inequalities $\theta_3 \ge \theta_2 \ge \theta_1$ are strict, that is, if $\theta_3 > \theta_2 > \theta_1$, then as $n_i \to \infty$, $i = 1, 2, 3$ the constrained and unconstrained estimators

Table 3.2. The relationship between the naive MLEs and the order-restricted MLEs.

Observed order of naive MLEs	Case	Ordering of constrained MLEs
$\theta_1^* < \theta_2^* < \theta_3^*$	I	$\hat{\theta}_1 < \hat{\theta}_2 < \hat{\theta}_3$
$\theta_1^* < \theta_3^* < \theta_2^*$	II	$\hat{\theta}_1 < \hat{\theta}_2 = \hat{\theta}_3$
$\theta_2^* < \theta_1^* < \theta_3^*$	III	$\hat{\theta}_1 = \hat{\theta}_2 < \hat{\theta}_3$
$\theta_2^* < \theta_3^* < \theta_1^*$	II or IV	$\hat{\theta}_1 = \hat{\theta}_2 < \hat{\theta}_3$ or $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}_3$
$\theta_3^* < \theta_1^* < \theta_2^*$	III or IV	$\hat{\theta}_1 < \hat{\theta}_2 = \hat{\theta}_3$ or $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}_3$
$\theta_3^* < \theta_2^* < \theta_1^*$	IV	$\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}_3$

agree with probability tending to one. Consequently the constrained estimators are nearly independent in large samples. Clearly if there are equalities among the parameters, the estimators will be dependent.

The ideas above can be implemented directly when estimating binomial proportions in $K > 3$-ordered populations. More interestingly the same ideas apply in the context of nonparametric estimation of two (or more) distribution functions. Suppose that $X_{i1},\ldots,$ $X_{in_i} \sim F_i$ for $i = 1,2$ and that it is known that the distribution functions are arbitrary but stochastically ordered, that is, $F_2(x) \geq F_1(x)$ for all $x \in \mathbb{R}$. Fix the value of x and note that

$$Y_i = \sum_{j=1}^{n_i} \mathbb{I}_{\{X_{ij} \leq x\}} \tag{3.18}$$

follows a $\text{Bin}(n_i, \theta_i)$ distribution where $\theta_i = F_i(x)$ and it follows that $\theta_2 \geq \theta_1$. Estimating the binomial parameters (θ_1, θ_2) under order restrictions is straightforward as indicated above. Varying the value of x we derive estimates for the distribution functions over their entire range. Pursuing the mathematics, we recover the estimates derived initially by Hogg [10] and discussed in depth by El Barmi and Mukerjee [5] and the references therein. We note that this estimator is not the nonparametric maximum likelihood estimator derived by Brunk et al. [3]. Let $F_i^*(x)$ and $\hat{F}_i(x)$ denote the naive and constrained estimators of F_i at x. Note that $F_i^*(x)$ is the well-known empirical distribution function. It follows that $\hat{F}_1(x) = F_1^*(x)$ and $\hat{F}_2(x) = F_2^*(x)$ whenever $F_2^*(x) \geq F_1^*(x)$, otherwise

$$\hat{F}_1(x) = \hat{F}_2(x) = \frac{n_1 F_1^*(x) + n_2 F_2^*(x)}{n_1 + n_2}. \tag{3.19}$$

A proof that $\hat{F}_i(x)$, $i = 1,2$ are distribution functions may be found in the appendix. Note that the resulting estimates are nothing but the point-wise isotonic regression of the unconstrained empirical distribution functions. For more on isotonic regression see Robertson et al. [17].

3.3. Bioequivalence. Two treatments are said to be equivalent if their mean responses are similar. The term *bioequivalence* is widely used in the pharmaceutical industry to describe different drug formulations with similar absorption characteristics. We will say

that treatments i and j are bioequivalent if $|\theta_i - \theta_j| \leq \Delta$, where θ_i denotes the mean response in group i, θ_j is similarly defined, and Δ is a prespecified, positive constant, describing our tolerance for differences among the means. This form of bioequivalence is known as *average bioequivalence*. For an in-depth statistical analysis of the bioequivalence problem see Berger and Hsu [1] and the references therein. The bioequivalence null hypothesis states that the differences between the treatment means are larger than Δ, that is, $H_0 : |\theta_i - \theta_j| > \Delta$. The alternative hypothesis is $H_1 : |\theta_i - \theta_j| \leq \Delta$. Thus rejecting the null implies bioequivalence. Estimating the parameters under both the null and the alternative is of great interest. Both are constrained estimation problems that may be solved using KT theory. We develop an estimation procedure under the alternative. A similar procedure applies under the null.

Consider the following simplified set up. Let $\overline{X}_i$ denote the sample average in the ith group. Assume that $\overline{X}_i$ all follow a normal distribution with equal variances, which we set, without loss of generality, equal to unity. Therefore the log-likelihood is

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{3}(\overline{x}_i - \theta_i)^2. \tag{3.20}$$

The bioequivalence hypothesis states that $|\theta_i - \theta_j| \leq \Delta$ for $1 \leq i, j \leq 3$. Clearly these constraints are not differentiable. However they may be equivalently rewritten as

$$\begin{aligned} &c_1(\theta) = \Delta - (\theta_1 - \theta_2), \qquad c_2(\theta) = (\theta_1 - \theta_2) + \Delta, \qquad c_3(\theta) = \Delta - (\theta_2 - \theta_3), \\ &c_4(\theta) = (\theta_2 - \theta_3) + \Delta, \qquad c_5(\theta) = \Delta - (\theta_1 - \theta_3), \qquad c_6(\theta) = (\theta_1 - \theta_3) + \Delta. \end{aligned} \tag{3.21}$$

Note that there are three pairs of constraint functions. Each pair of constraints corresponds to one of the original equivalence relations. In order to maximize the log-likelihood on the feasible set we differentiate the Lagrangian and set the resulting equations equal to zero. Thus we solve

$$\nabla L = \begin{pmatrix} \overline{x}_1 - \theta_1 - \lambda_1 + \lambda_2 - \lambda_5 + \lambda_6 \\ \overline{x}_2 - \theta_2 + \lambda_1 - \lambda_2 - \lambda_3 + \lambda_4 \\ \overline{x}_3 - \theta_3 + \lambda_3 - \lambda_4 + \lambda_5 - \lambda_6 \end{pmatrix} = 0 \tag{3.22}$$

together with

$$c_i(\theta) \geq 0, \quad \lambda_i \geq 0, \quad \lambda_i c_i(\theta) = 0 \quad \text{for } i = 1, \ldots, 6. \tag{3.23}$$

Obviously, the solution is determined by which constraints are effective at the optimum. In principle, a complete solution of (3.22) and (3.23) requires the consideration of all possible combinations of effective constraints. Enumeration shows that there are (potentially) 2^6 such possibilities. However a careful analysis shows that the true number of possibilities is much smaller.

Without loss of generality, relable the treatments in such a way that $\overline{x}_1 > \overline{x}_2 > \overline{x}_3$. Clearly this ordering of the observed data induces the same ordering for the estimated

Table 3.3. Solutions for the constrained estimation problem involving three bioequivalent means.

	Case I	Case II	Case III	Case IV	Case V
$\hat{\theta}_1$	x_1	$\frac{x_1+x_3+\Delta}{2}$	$\frac{x_1+x_2+x_3+2\Delta}{3}$	$\frac{x_1+x_2+x_3+\Delta}{3}$	$\frac{x_1+x_2+x_3+\Delta}{3}$
$\hat{\theta}_1$	x_2	x_2	$\frac{x_1+x_2+x_3-\Delta}{3}$	$\frac{x_1+x_2+x_3+\Delta}{3}$	$\frac{x_1+x_2+x_3}{3}$
$\hat{\theta}_1$	x_3	$\frac{x_1+x_3-\Delta}{2}$	$\frac{x_1+x_2+x_3-\Delta}{3}$	$\frac{x_1+x_2+x_3-2\Delta}{3}$	$\frac{x_1+x_2+x_3-\Delta}{3}$
$\hat{\lambda}_1$	0	0	$\frac{x_1-2x_2+x_3-\Delta}{3}$	0	$\frac{2x_1-x_2-x_3-\Delta}{3}$
$\hat{\lambda}_3$	0	0	0	$\frac{-x_1+2x_2-x_3-\Delta}{3}$	$\frac{x_1+x_2-2x_3-\Delta}{3}$
$\hat{\lambda}_5$	0	$\frac{x_1-x_3-\Delta}{2}$	$\frac{x_1+x_2-2x_3-\Delta}{3}$	$\frac{2x_1-x_2-x_3-\Delta}{3}$	0

means. The differences $\overline{x}_i - \overline{x}_j$ for $i < j$ are always positive. Therefore after relabelling, the constraints c_2, c_4, and c_6 hold automatically. Applying the principle of complementary slackness, we set $\widehat{\lambda}_2 = \widehat{\lambda}_4 = \widehat{\lambda}_6 = 0$. Thus only combinations of the constraints c_1, c_3, and c_5 need be considered. There are 2^3 possible combinations of these constraints that can, in principle, be effective at the optimum. These are $\{\varnothing\}$, $\{c_1\}$, $\{c_3\}$, $\{c_5\}$, $\{c_1,c_3\}$, $\{c_1,c_5\}$, $\{c_3,c_5\}$, and $\{c_1,c_3,c_5\}$. By construction $\overline{x}_1 - \overline{x}_3 > \overline{x}_1 - \overline{x}_2$ therefore if c_1 is effective, c_5 must also be. Similarly, if the constraint c_3 is effective, then c_5 must be. Moreover it is easy to check that the three constraints c_1, c_3, and c_5 are not jointly compatible but all pairs are. Therefore if c_1 and c_3 are effective, then c_5 is automatically ineffective. Hence only five solutions are possible; these are summarized in Table 3.3: see [4].

In addition

$$\nabla c_1(\theta) = (-1,1,0), \qquad \nabla c_3(\theta) = (0,-1,1), \qquad \nabla c_5(\theta) = (-1,0,1). \tag{3.24}$$

It follows that the constraint qualification holds for all possible combinations of constraints which can be effective at the optimum. Therefore the conditions of KT are satisfied. Moreover it is easily verified that these are global maxima. Extensions to more than three treatments are clear. It is worth noting that typically, even in multivariate bioequivalence problems, treatments are compared two at a time. Our derivations point the way for estimation and testing procedures which consider simultaneous bioequivalence for large number of treatments. Further research on inferential procedures for this model are warranted.

4. Summary and discussion

We introduce the theorem of KT and describe how it applies in three very different constrained estimation problems. In our examples the objective function is the log-likelihood and our estimators are MLEs. The method, however, is clearly applicable in more general settings and to other types of estimating equations. In our examples KT finds the

global maximum. This remains true in many statistical problems because the objective functions are often concave and the constraints define a convex (even bounded) region. Although the models in Section 3 are well known and had been analyzed using different approaches, our derivations add a unique perspective. For example, in the case of ridge regression we explicitly relate the dual parameters λ and K. Next, estimators for ordered (event) probabilities under binomial sampling, which are of intrinsic interest, are used to derive estimators for the empirical distributions function. Finally our treatment of the bioequivalence problem extends the usual analysis and shows how to generalize to an arbitrary number of treatments. Note that explicit expressions for the MLEs are obtained in all three cases. This is not always true even when the constraints are linear. Consider, for example, the regression problem with positivity constraints (i.e., $\theta \geq 0$ component-wise). As noted by Wang et al. [20] a constrained linear model can be solved using the simplex method in small number of steps. However constrained estimation in generalized linear models is more complicated because the objective function is nonlinear. More powerful tools need to be used in conjunction with KT to find a solution in such situations. Finally we would like to mention the papers Dykstra and Wollan [4] who introduce a partial iterated KT theorem for problems with large number of constraints. Such methods seem applicable, for example, in the evaluation of bioequivalence of a large number of treatments.

Appendix

In Section 3.2 we derived estimators for the distribution functions under the assumption that $F_2(x) \geq F_1(x)$ for all $x \in \mathbb{R}$. In particular we showed that

$$\left(\widehat{F}_1(x), \widehat{F}_2(x)\right) = \begin{cases} \left(F_1^*(x), F_2^*(x)\right) & \text{if } F_2^*(x) \geq F_1^*(x), \\ \left(\dfrac{n_1F_1^*(x) + n_2F_2^*(x)}{n_1+n_2}, \dfrac{n_1F_1^*(x) + n_2F_2^*(x)}{n_1+n_2}\right) & \text{if } F_2^*(x) < F_1^*(x), \end{cases} \tag{A.1}$$

where $F_i^*(x)$ for $i = 1,2$ are the empirical distribution functions. By construction $\widehat{F}_1(x) \leq \widehat{F}_2(x)$ for all x.

Proposition A.1. *The functions $\widehat{F}_i(x)$ defined in (A.1) are proper distribution functions.*

Proof. We divide the proof into three parts. (1) Let

$$\begin{aligned} m &= \min\{X_{ij} \mid i = 1,2,\ j = 1,\ldots,n_i\}, \\ M &= \max\{X_{ij} \mid i = 1,2,\ j = 1,\ldots,n_i\}. \end{aligned} \tag{A.2}$$

Clearly $F_1^*(x) = F_2^*(x) = 0$ for all $x < m$ and $F_1^*(x) = F_2^*(x) = 1$ for all $x > M$. Substituting in (A.1) we find that $\widehat{F}_1(x) = \widehat{F}_2(x) = 0$ for all $x < m$ and that $\widehat{F}_1(x) = \widehat{F}_2(x) = 1$ for all $x > M$. Consequently

$$\lim_{x \longrightarrow -\infty} \widehat{F}_i(x) = 0, \quad \lim_{x \longrightarrow \infty} \widehat{F}_i(x) = 1 \quad \text{for } i = 1,2. \tag{A.3}$$

(2) Let $s < t$. By definition we have

$$F_1^*(s) \le F_1^*(t), \quad F_2^*(s) \le F_2^*(t). \tag{A.4}$$

In addition only one of the four possible events may occur, either (i) $F_1^*(s) \le F_2^*(s)$ and $F_1^*(t) \le F_2^*(t)$; or (ii) $F_1^*(s) \le F_2^*(s)$ and $F_1^*(t) > F_2^*(t)$; or (iii) $F_1^*(s) > F_2^*(s)$ and $F_1^*(t) \le F_2^*(t)$; or (iv) $F_1^*(s) > F_2^*(s)$ and $F_1^*(t) > F_2^*(t)$. It is easily verified that (i) and (A.4) imply that

$$\hat{F}_i(s) = F_i^*(s) \le F_i^*(t) = \hat{F}_i(t). \tag{A.5}$$

Condition (ii) and (A.4) imply that

$$\begin{aligned} \hat{F}_i(s) = F_i^*(s) \le F_2^*(s) &= \frac{n_1 F_2^*(s) + n_2 F_2^*(s)}{n_1 + n_2} \\ &\le \frac{n_1 F_2^*(t) + n_2 F_2^*(t)}{n_1 + n_2} \le \frac{n_1 F_1^*(t) + n_2 F_2^*(t)}{n_1 + n_2} = \hat{F}_i(t). \end{aligned} \tag{A.6}$$

Condition (iii) and (A.4) imply that

$$\hat{F}_i(s) = \frac{n_1 F_1^*(s) + n_2 F_2^*(s)}{n_1 + n_2} \le \frac{n_1 F_1^*(s) + n_2 F_1^*(s)}{n_1 + n_2} = F_1^*(s) \le F_1^*(t) \le F_i^*(t) = \hat{F}_i(t). \tag{A.7}$$

Condition (iv) and (A.4) imply that

$$\hat{F}_i(s) = \frac{n_1 F_1^*(s) + n_2 F_1^*(s)}{n_1 + n_2} \le \frac{n_1 F_1^*(t) + n_2 F_1^*(t)}{n_1 + n_2} = \hat{F}_i(t). \tag{A.8}$$

We conclude that

$$\hat{F}_i(s) \le \hat{F}_i(t) \quad \text{for } i = 1, 2. \tag{A.9}$$

(3) The functions $F_i^*(x)$ are right continuous and therefore so are their linear combinations, maximums, and minimums. It immediately follows that

$$\lim_{x \downarrow x_0} \hat{F}_i(x) = \hat{F}_i(x_0) \quad \text{for } i = 1, 2. \tag{A.10}$$

Hence the constrained estimators satisfy (A.4), (A.8), and (A.10), the defining properties of distribution functions. □

References

[1] R. L. Berger and J. C. Hsu, *Bioequivalence trials, intersection-union tests and equivalence confidence sets*, Statistical Science **11** (1996), no. 4, 283–319.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[3] H. D. Brunk, W. E. Franck, D. L. Hanson, and R. V. Hogg, *Maximum likelihood estimation of the distributions of two stochastically ordered random variables*, Journal of the American Statistical Association **61** (1966), 1067–1080.
[4] R. L. Dykstra and P. C. Wollan, *Constrained optimization using iterated partial Kuhn-Tucker vectors*, Reliability and Quality Control (Columbia, Mo, 1984), North-Holland, Amsterdam, 1986, pp. 133–139.
[5] H. El Barmi and H. Mukerjee, *Inferences under a stochastic ordering constraint: the k-sample case*, Journal of the American Statistical Association **100** (2005), no. 469, 252–261.
[6] C. J. Geyer and E. A. Thompson, *Constrained Monte Carlo maximum likelihood for dependent data. With discussion and a reply by the authors*, Journal of the Royal Statistical Society. Series B **54** (1992), no. 3, 657–699.
[7] M. H. J. Gruber, *Improving Efficiency by Shrinkage. The James-Stein and Ridge Regression Estimators*, Statistics: Textbooks and Monographs, vol. 156, Marcel Dekker, New York, 1998.
[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, 2001.
[9] A. E. Hoerl and R. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics **12** (1970), 55–67.
[10] R. V. Hogg, *On models and hypotheses with restricted alternatives*, Journal of the American Statistical Association **60** (1965), 1153–1162.
[11] M. Jamshidian and P. M. Bentler, *A modified Newton method for constrained estimation in covariance structure analysis*, Computational Statistics & Data Analysis **15** (1993), no. 2, 133–146.
[12] K. Lange, *Numerical Analysis for Statisticians*, Statistics and Computing, Springer, New York, 1999.
[13] S. Y. Lee, *Constrained estimation in covariance structure analysis*, Biometrika **66** (1979), no. 3, 539–545.
[14] S.-Y. Lee, *The multiplier method in constrained estimation of covariance structure models*, Journal of Statistical Computation and Simulation **12** (1981), 247–257.
[15] C. K. Liew, *Inequality constrained least-squares estimation*, Journal of the American Statistical Association **71** (1976), no. 355, 746–751.
[16] C. P. Robert and J. T. G. Hwang, *Maximum likelihood estimation under order restrictions by the prior feedback method*, Journal of the American Statistical Association **91** (1996), no. 433, 167–172.
[17] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order Restricted Statistical Inference*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Chichester, 1988.
[18] M. J. Silvapulle and P. K. Sen, *Constrained Statistical Inference*, Wiley Series in Probability and Statistics, John Wiley & Sons, New Jersey, 2005.
[19] R. K. Sundaram, *A First Course in Optimization Theory*, Cambridge University Press, Cambridge, 1996.
[20] D. Q. Wang, S. Chukova, and C. D. Lai, *On the relationship between regression analysis and mathematical programming*, Journal of Applied Mathematics & Decision Sciences **8** (2004), no. 2, 131–140.

Ori Davidov: Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905, Israel
E-mail address: davidov@stat.haifa.ac.il