Информационные процессы, Том 18, № 4, 2018, стр. 261–295
© 2018 Malyutov.

REAGAN Estimates of a Quasilinear Regression

M. Malyutov*

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
E-mail: m.malioutov@neu.edu

Received 03.09.2018

Abstract—We study Design and Analysis of non-linear in parameters smooth Quasilinear Regression Model (abbreviated as Q-model) which admits in every iteration almost the same type of analysis as linear regression for **simultaneous iterative estimation of both the mean and variance** of observations. Q-models are more flexible in applications than Linear models. We prove local convergence of iterations $\theta^s, s \in Z$, in probability and the asymptotic normality of $\hat{\theta} = \lim_{s\to\infty} \theta^s$ in iteration procedure called **REweighted Algorithm of GAuss-Newton** and illustrate its broad use by many appropriate applications.

KEYWORDS: Gauss-Newton algorithm, iterations, convergence in probability, asymptotic normality, maximum likelihood, exponential families,

1. INTRODUCTION: MAIN RESULTS AND APPLICATIONS

Design and Analysis in *Linear models* is a favorite in Statistics with the mean of unknown uni/multivariate distribution estimated from empirical data while the **covariances are known** up to a scalar multiplier. The Least Squares (**LSE**) method of Gauss estimates the mean of multivariate distribution and a scalar multiplier of Covariance from empirical data. Algebraic formulas for estimates and their statistical characteristics make LSE most transparent and attractive.

A general regression of a parametric distribution family P_θ is the **best approximation of the response via explanatory variables** in square risk sense. Usually, the covariance $D(\theta)$ of this prediction **depends on parameters** θ of the distribution and the regression cannot be described by a Linear model. If neither theory, nor the data amount do not allow use of distribution P_θ, but second moments are available, then the nonparametric approach is justified which we start describing.

We survey more broadly applicable and flexible Quasilinear Multivariate Regression Model Q that has been implicitly used in various applications since 1950's. It is based on the following basic assumptions:

(i) The mean and Covariance matrices $D_i(\theta)$ of independent observations $y_i \in E^m, m \in Z$, are smooth functions of the **same unknown parameter** $\theta \in \Theta \subset E^p$.

(ii) **Identifiability**: Parameter θ is uniquely determined by vector $(\mu_1(\theta), \dots, \mu_N(\theta))$.

We call this model **Q-Model**. Non-smooth in θ models (say, change point and range of distribution estimation)) require alternative types of analysis.

Remark. *Condition (ii) distinguishes Q-models from the Variance Components (VC) model* (fitted in our section 12 using quadratic in $\mathbf{y}$ updates), where covariance coefficients VC are separate parameters apart from the mean.

Algebraic formulas for estimates of θ in Q-model are not available. We prove local convergence *in probability* of iterations $\theta^s, s \in Z$ and asymptotic normality (**AN**) of $\hat{\theta} = \lim_{s\to\infty} \theta^s$ in iteration procedure which generalizes the classical Gauss-Newton algorithm in the following aspect: the difference $\theta^{s+1} - \theta^s$ is the **weighted** LSE of linearized at θ^s model and plugged in weights are given by the formula $V_i = D_i^{-1}(\theta^s)$. Let us call this algorithm **Reweighted Algorithm of Gauss-Newton**, or **REAGAN**.

Remark. The Gauss modification of the Newton Algorithm was invented for simplifying the latter in analyzing **random nonlinear regression data** by replacing numerically cumbersome inverse of the Hessian with the inverse of the easily computable **mean Hessian**. An additional advantage is **local convergence of iterations** which is not guaranteed even for deterministic applications (say $\sqrt{\cdot}$) of the Newton algorithm.

In section **5**, we prove: if initial approximation (starting point) θ^0 is a consistent estimate of θ^*, then estimate $\hat{\theta}$ is an AN estimate of θ^* with asymptotic covariance $(\sqrt{N}(\partial F(\theta^*)^T D(\theta^*)(\partial F(\theta^*))^{-1}$.

Milder conditions for initial guess $\hat{\theta}_0$ are established in section **5** by arguments of the *type two embeddings* of Banach spaces.

Previously, in section **4**, we prove the asymptotic normality (AN) and convergence of moments $\sqrt{N}(\theta^V - \theta^*)$, where θ^V is the LSE with a **fixed** weight function $V(x)$. This implies $\sqrt{N}$-consistency of θ^V; we prove convergence of moments $\sqrt{N}(\theta^1 - \theta^*)$ in section **5**. In section **6**, we prove the **Local Asymptotic Minimaxity** (**LAM**) of **REAGAN** for Q-models in the class of linear in y updates to an initial guess.

In section **10**, we use the following observation of Generalized Linear models [37]: estimate $\hat{\theta}$ satisfies maximum likelihood equation, if multivariate distribution of measurements belongs to a *Regular Exponential Family* and prove the minimaxity of **LAM**-risk of **REAGAN** with respect to any statistics and distribution families with the same means $\mu_i(\theta)$ and $Cov_\theta[y_i] \leq D_i(\theta)$.

In section **9**, we introduce a polynomial generalization: F^k-model. In particular, F^2 is a natural generalization of the **Variance Component model** (**VCm**). We prove: **REAGAN** has the **LAM** property with respect to the class of polynomial functions in y_i of degree k for the extended **F**-model of larger dimension introduced there.

Next section **12** deals with a particular case of F^k-model: one-way Gaussian Variance Components. We prove the Local Asymptotic Normality of estimates and establish unusual properties of optimal design.

A collection of other applied examples reduced to Q-models is reviewed in sections **11, 13–15**, including:

1. a multivariate regression with unknown constant covariance matrix of measurements (section **11**);
2. multisample mixture estimation (section **13**);

   We only sketch several other applications:
3. the popular **M**-estimation can be also represented as a particular case of **REAGAN** algorithm.
4. density parameter estimation from stratified sample (section **14**);
5. the first approximation for regression estimates with small random error in controlled variables parameters (section **15**).

Rather distant interesting application of **REAGAN** for Q-models is in *spectral domain* of Gaussian multivariate time series [19] which is beyond the scope of this survey as well as **REAGAN** use for normalization of RNA microarrays via orthogonal regression [34].

Non-linear regression models were popular publication topics until mid-eighties of twentieth century. Strangely enough, a very special case of Q-models called Generalized Linear models [37] was more popular than broader Q-models. Later, non-parametric models replaced them as the main focus of statistical community. Interest to non-linear parametric models seems to get more ground now. In particular, design and analysis for very many Q-models with their application to **clinical trials** are studied in [8], where some of our results are reproduced.

2. NOTATION AND ABBREVIATIONS

$\doteq, \exists$ mean *equality by definition* and 'there exists',

s.t. abbreviates *'such that'*,

w.r.t. means 'with respect to';

$\mathbf{R}^k$ $(\mathbf{E}^k)$ is the k-dimensional (*Euclidean*) linear vector space (with norm $\|\cdot\|$);

$\mathbf{N}, (\mathbf{Z})$ are the sets of natural numbers (integers),

$\lceil x \rceil = \min\{z \in \mathbf{Z} : z \geq x\}$;

$\mathbf{I_A}$ is the indicator of the set $\mathbf{A}$,

$Int[\mathbf{A}]$ is the interior of an open set $\mathbf{A}$;

$\mathbf{B}_\varepsilon$ is a ball with center $\theta^* \in \mathbf{E}^p$ and radius ε for a fixed $\theta^* \in Int[\mathbf{A}]$;

For $F : \mathbf{R}^k \times \mathbf{R}^m \to \mathbf{R}$, $\Delta\mathbf{F}(x, \theta, \theta^*) = \mathbf{F}(x, \theta) - \mathbf{F}(x, \theta^*)$;

F^* is the value of the function $F(\theta)$ at $\theta = \theta^*$;

If $F(x, \theta)$ is continuously differentiable and depends also on chance (random), then the Lagrange form of its first order Taylor expansion is preferable to the conventional form since it enables proof that $\Delta\mathbf{F}(x, \theta, \theta^*)$ is measurable, integrable, etc.

$$\Delta\mathbf{F}(\theta, \theta^*) = \int_0^1 (\partial\mathbf{F}(\theta + \lambda(\theta^* - \theta))/\partial\theta) d\lambda \cdot (\theta - \theta^*). \tag{1}$$

The corresponding Lagrange form of the second order Taylor expansion (12.4.3) is used in section 12 for statistical analysis of quadratic approximations.

$\det \mathbf{A}$, $\mathbf{A}^T$, $tr\mathbf{A}$ are determinant, transpose and trace of matrix $\mathbf{A}$; $Vec[\mathbf{W}(x_1^N)] := (\mathbf{W}^T(x_1), \ldots, \mathbf{W}^T(x_N))^T$; $\mathbf{A}^-$ is the generalized Moore-Penrose inverse of matrix $\mathbf{A}$; $\mathbf{I}$ is the identity matrix; $\mathbf{A} \geq \mathbf{B}$ $(\mathbf{A} > \mathbf{B})$, if $\mathbf{A} - \mathbf{B}$ is symmetric and nonnegative (positive) definite; $diag\{\lambda_1^n\}$ is a block-diagonal matrix with blocks $\lambda_1, \ldots, \lambda_n$; $\mathbf{1}_k \in \mathbf{E}^k = (1, \ldots, 1)^T$;

Borel subsets $\mathbf{A}_N$ of $\mathbf{R}^k$ hold in probability, if their lower measure $\underline{P}(\mathbf{A}_N) \to 1$ as $N \to \infty$, i.e. $\exists$ a sequence of open sets $B_{mN} \subseteq A_N$ s.t. ; $P(\mathbf{B}_{mN}) \to 1$; $\xi_N \Rightarrow \mathbf{N}(a, D)$ denotes the weak convergence of random vector ξ_N to the Normal distribution $\mathbf{N}(a, D)$ with mean a and covariance matrix D.

2.1. Assumptions

Let $P_{x,\theta}$ be a family of distributions on Borel subsets of $\mathbf{R}^m$ or on finite set of elements of $\mathbf{R}^m, m \in \mathbf{N}$, which depend on known $x_i \in X$ and $\theta \in \Theta \subset \mathbf{E}^p, p \in \mathbf{N}$, where $\Theta \subset \mathbf{E}^p$ is a compact set.

We observe a sequence of independent random variables

$$y = y^{(N)} = Vec\{y_1^N\}, y_i = y_i^{(N)} \in \mathbf{E}^m,$$

$P(y_i \in B) = P_{x_i,\theta^*}(B)$ for some unknown $\theta^* \in Int\Theta$ and any Borel sets $B \subset \mathbf{R}^m$.

$\mathbf{E}_\theta[\xi]$ and $\mathbf{Cov}_\theta[\xi]$ are expectation and covariance matrix of random vector ξ with distribution $\mathbf{P}_\theta$

We use the following assumptions:

(A1.a): *Mean* $\mathbf{E}_\theta[y_i] = \mu_i(\theta) = \mu(x_i, \theta)$, *where* $\mu(x, \theta)$ *is a bounded smooth function such that*

(A1.b): $\varphi(x, \theta) = \partial\mu(x, \theta)/\partial\theta$ *is a continuous bounded* $(m \times p)$*-matrix function on* $X \times \Theta$

(A1.c): *The same requirements for* $\partial\varphi/\partial\theta_i, i = 1, \ldots, p$

(A2): *There exists a covariance matrix* $\mathbf{Cov}_\theta[y_i] = D(x_i, \theta)$ *such that* $\partial D(x, \theta)/\partial\theta$ *is a smooth bounded function on* $X \times \Theta$ *with*

$$\inf_{X\times\Theta} \det D(x, \theta) > 0 \tag{2}$$

(A3.a): *Distributions* ε_N *on* X: $\varepsilon_N(x) = N^{-1}\sum_{i=1}^{N} \mathbf{I}_x(x_i)$ *weakly converge to a probability measure* ε *and*

(A3.b):

$$\int \varphi^T(x, \theta) D^{-1}(x, \theta)\varphi(x, \theta) d\varepsilon(x) \doteq m(\theta) \tag{3}$$

satisfies condition

$$\inf_{\Theta} \det m(\theta) > 0. \tag{4}$$

(A3.c): $\forall\ \theta, \theta' \in \Theta, \theta \neq \theta'$ *and continuous bounded weight function* $V(x) > 0$ *such that*

$$0 < \inf \det V(x) \leq \sup \det V(x) < \infty$$

we have that

$$R_V(\theta, \theta') \doteq \int \Delta\mu^T(x, \theta, \theta')V(x)\Delta\mu(x, \theta, \theta')d\varepsilon(x) > 0. \tag{5}$$

Remark 1. (1) Our results are easily generalized for a bounded number of blocks of measurements B_i and $N \to \infty$ in each of them, (A1)-(A3) hold for the same Θ and some $m_i, D_i, \mu_i, \varepsilon_{N(i)}$.

(2) Simultaneous measurable w.r.t. x diagonalization of two weight matrices $V(x)$ and $V'(x)$ of the quadratic form (5) yields: if (A3.c) holds for some family of matrices $V(x)$, then it holds for all families of matrices with the same properties as in (A3.c).

3. AUXILLARY RESULTS

The following development follows the functional approach to asymptotic statistics originated in the proof of limiting distribution for the Kolmogorov (1933) goodness of fit non-parametric test.

Theorems and formulas preceding section 12 are enumerated without index of section. Theorem 1 is a correct replacement of erroneous section 4.3.8 in [47] (its author was apparently unaware of the functional approach). Section 4.3.8 **implied unjustifiably mild regularity conditions** for many convergence theorems in his fundamental textbook [47] and in numerous other books and papers that cited [47] (see e.g.[5, 6]).

The error was caused by replacing $P(\sup A(\theta) < \varepsilon)$ with $\sup P(A(\theta) < \varepsilon)$ for a family $A(\theta)$ of random variables. The counterexamples to the statement of [47], section 4.3.8, are in [30].

Theorem 1. *Suppose a sequence of measurable functions* $g_N(x, \theta, y)$ *is uniformly continuous on* $X \times \Theta \times \mathbf{E}^m$ *and converges uniformly to* $g(x, \theta, y)$ *on* $X \times \Theta$ *a.e. with respect to y-distribution, and* $\forall x, \theta \in X \times \Theta \quad |g_N(x, \theta, y)| \leq \psi(y)$, *where* $M = \sup_X E_{x,\theta^*}\psi^2(y) < \infty$. *Let* $G_N(x, \theta) = E_{x,\theta^*} g_N(x, \theta, y)$ *be continuous on* $X \times \Theta$ *uniformly w.r.t.* N *and assumption (A3.a) is satisfied. Then for*

$$S_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} g_N(x_i, \theta, y_i)$$

the following is true:

(1.a) $S_N(\theta) \to \bar{G}(\theta) \doteq \int G d\varepsilon \quad (G(x, \theta) = \lim_{N\to\infty} G_N(x, \theta))$ *in probability uniformly in* $\theta \in \Theta$.

(1.b) If g_N *does not depend on* y, *then 'in probability' in (1.a) can be omitted.*

(2) $\lim_{r\to 0}\overline{\lim}_{N\to\infty} P(\sup_{B_r}\|S_N(\theta)-\bar{G}(\theta)\| > \kappa) = 0 \;\; \forall\, \kappa > 0.$

(3) If $\hat{\theta}_N \to \theta^*$ *in probability, then* $S_N(\hat{\theta}_N) \to \bar{G}(\theta^*)$ *in probability.*

Proof. We have

$$\sup_{X\times\Theta} |G_N(x,\theta)| \le L \sup_X E_{x,\theta^*}\psi(y) < \infty;$$

$$S_N(\theta)-\bar{G}(\theta) = \int \xi_N(x,\theta,y)d\varepsilon_N + \int G_N(d\varepsilon_N - d\varepsilon) + \int (G_N - G)d\varepsilon,$$

$$where \quad \xi_N(x_i,\theta,y_i) = g_N(x_i,\theta,y_i) - G_N(x_i,\theta).$$

From (A3.a) and Prokhorov's Theorem, we have:

$$\forall \;\; \lambda > 0 \;\; \exists \; compact \; K_\lambda \subset X \;\; s.\, t. \;\; \sup_{N\in\mathbf{N}} \varepsilon_N(K_\lambda) > 1 - \frac{\lambda}{8L}.$$

The uniform continuity of G_N on $K_\lambda \times \Theta$ implies that

$$\exists\; \kappa = \kappa(\lambda) \; s.t. \; \sup\{\left|\Delta G_N(x,\theta,\theta')\right| \; : \; \theta,\theta' \in \Theta, \; \left\|\theta-\theta'\right\| < \kappa\} < \frac{\lambda}{8} \quad \forall\, N \in \mathbf{N}.$$

We have for $\forall N > N^*(\lambda)$:

$$\sup_\Theta \left|\int (G_N - G)d\varepsilon\right| < \frac{\lambda}{4}, \;\; \sup_{A_\kappa}\left|\int G_N(d\varepsilon_N - d\varepsilon)\right| < \frac{\lambda}{4},$$

where A_κ is a finite κ-net on Θ. Then

$$\sup_\Theta \left|\int G_N d\varepsilon_N - \bar{G}\right| < \lambda \;\; for \; N > N^*(\lambda),$$

i.e. (1.b) follows. Further,

$$\sup_\Theta P(\left|\int \xi_N d\varepsilon_N\right| \ge \nu) \le \frac{\sup_\Theta E(\int \xi_N d\varepsilon_N)^2}{\nu^2} \le \frac{M}{N\nu^2} \to 0 \;\; as \;\; N \to \infty,$$

which implies (1.a).

If for $\forall\; \lambda > 0$, $\omega_N(\gamma,y,\lambda)$ is the continuity modulus of $\xi_N(x,\theta,y)$ on the compact $K_{\lambda/2}\times\Theta$, then

$$P(\sup_{B_\gamma}\left|\int \Delta\xi_N(x,\theta,\theta^*,y)d\varepsilon_N\right| > \frac{\kappa}{2}) \le 2\frac{E\sup_{B_\gamma}\left|\int \Delta\xi_N d\varepsilon_N\right|}{\kappa},$$

$$E\sup_{B_\gamma}\left|\int \Delta\xi_N d\varepsilon_N\right| \le \int_{K_{\lambda/2}} E\sup_{B_\gamma}|\Delta\xi_N(x,\theta,\theta^*,y)|\, d\varepsilon_N + \frac{\lambda}{3} \le E\omega_N(\gamma,y,\lambda) + \frac{\lambda}{3} < \lambda,$$

when $\gamma < \gamma^*(\lambda)$ by the Lebesgue Theorem. This and (1.b) imply (2) and consequently (3).

Remark 2. In the proof of Theorem **1** we used only the null correlation between $y_1,\ldots,y_N$.

Corollary 1. *Assuming (A1)-(A3) and result (1.b) of Theorem* **1**, *we have:*

$$m_N(\theta) \doteq \int \varphi^T(x,\theta)D^{-1}(x,\theta)\varphi(x,\theta)d\varepsilon_N(x) \to m(\theta)$$

uniformly in $\theta\in\Theta$, *and* $m(\theta)$ *is a continuous function of* θ.

As an example, let us consider a Multinomial model. Let y_i be m-vector s. t. its only nonzero component $y_{ij} = 1$ means that a particle in the ith experiment gets into jth urn, $j = 1, \ldots, m$. Then

$$Cov_\theta[y_i] = diag\{p_1^m(\theta)\} - p(\theta)p^T(\theta)\}, \qquad \det Cov_\theta[y_i] \equiv 0$$

with $p^T(\theta) = (p_1(\theta), \ldots, p_m(\theta))^T$ is the vector of probabilities for particles to get into urns satisfying $\sum p_i(\theta) \equiv 1$. We can ensure that assumption (2) holds by eliminating the first component of vectors y_i. Instead, to avoid asymmetry and complex calculations, we can get rid of (2) and replace (3) with

$$\int \varphi^T(x,\theta)D^-(x,\theta)\varphi(x,\theta)d\varepsilon(x) = m(\theta), \tag{6}$$

where $D^-(x,\theta)$ is a continuous bounded on $X \times \Theta$ Moore-Penrose generalized inverse of $D(x,\theta)$ [3] satisfying (4). For Multinomial Models, basic models of the Analysis of Variance and other cases where y_i's have constant rank, it is possible to construct continuous generalized inverse of $D(x,\theta)$. In particular, for Multinomial models there is an obvious choice of a generalized inverse

$$D^-(x,\theta) = diag\{[p_1^m]^{-1}(\theta)\}$$

The above choice helps to construct the Q-model-based estimation in our sections 13-14.

4. LSE UNDER A CONSTANT KNOWN COVARIANCE MATRIX

Let us fix a matrix function $V(x) > 0$ as in (A3.c) and define LSE statistic:

$$\theta^V \doteq \arg\min_{\Theta} \sum \delta_i^T(\theta)V(x_i)\delta_i(\theta), \qquad \delta_i(\theta) \doteq y_i - \mu_i(\theta). \tag{7}$$

Assuming (A1.a), (A3.a) and (A3.c), θ^V is a correctly defined statistic for sufficiently large N (because the minimum in (7) is unique) and consistent estimator of θ^*. Let us give a simple proof of asymptotic normality of θ^V utilizing (A1.c). Let $Q_N(\theta)$ be a sequence of statistics under arg min sign in (7) considered as random fields of argument $\theta \in \mathbf{E}^p$. Introduce $R_N(t) \doteq Q_N(\theta) - Q_N(\theta^*)$, where $t = \sqrt{N}(\theta - \theta^*)$. Under assumptions (A3.a,b,c) and additional assumption:

(A4.(r)): for $\forall r > p$ *the moment of order* r *of distribution* $P_{x,\theta}$ *is bounded on* $X \times \Theta$

the following Theorem is true.

Theorem 2. *Random fields* $R_N(t)$ *with* t *belonging to some compact set* T *converge weakly to the Gaussian Random Field* $\{t^TAt - 2t^T\xi\}$, *where*

$$A = \int \varphi^{*T}V\varphi^*d\varepsilon,$$

$$\xi \sim \mathbf{N}(0,B), \quad B = \lim_{N\to\infty}\frac{1}{N}JDJ^T, \quad D = diag\{D(x_1^N,\theta)\}, \quad J = Vec\{\varphi^TV(x_1^N,\theta)\}.$$

Proof. Using (1), we get:

$$R_N(t) = \mathbf{I}_N(t) + \mathbf{II}_N(t),$$

$$\mathbf{I}_N(t) \doteq t^T\int\int d\lambda d\mu\frac{1}{N}\sum_{i=1}^N \varphi^T(x_i,\theta^* + \lambda\frac{t}{\sqrt{N}})V(x_i)\varphi(x_i,\theta^* + \mu\frac{t}{\sqrt{N}}),$$

$$\mathbf{II}_N(t) \doteq -2t^T\int_0^1 \frac{1}{\sqrt{N}}\sum_{i=1}^N \varphi^T(x_i,\theta^* + \lambda\frac{t}{\sqrt{N}})V(x_i)\delta_i^*d\lambda.$$

Theorem **1**, item (1.b) implies: expression $\mathbf{I}_N$ converges uniformly in $t \in T$ to t^TAt. The Lindeberg's condition for integrand in $\mathbf{II}_N$ holds uniformly in $t \in T$, which implies that $\mathbf{II}_N$ converges uniformly to the random field $-2t^T\xi$ on T.

Under conditions of Theorem **2**, the following Lemma is true.

Lemma 1. *There exist constants ρ_0, N_0, C such that for all $N > N_0$, $\rho > \rho_0$ the following inequality holds*

$$P\{\sqrt{N}\,\|\theta^V - \theta^*\| > \rho\} \le C\rho^{-r} \tag{8}$$

This result was proved in [16] for the univariate case of $\theta \in \mathbf{E}^1$ and generalized later by the same author to the case $p \ge 1$. The outline of the proof is as follows. The event in (8) is equivalent to the event

$$\sup_{\|t\|>\rho} (\mathbf{II}_N(t)/\mathbf{I}_N(t)) \ge 1. \tag{9}$$

The upper bound of probability in (9) is based, in particular, on a uniform in N bound of continuity modulus of random fields $\mathbf{I}_N(t)$ and $\mathbf{II}_N(t)$ which, in turn, is obtained using the Whittle inequality for the moments of linear form of independent random variables [46]. The same bounds also imply weak compactness of the family of measures corresponding to $R_N(t)$.

The associated with random field $-2t^T\xi$ probability measure in the space of continuous functions on T satisfies: measures corresponding to $\mathbf{II}_N(t)$ converge weakly to it. The functional $\arg\min R(t)$ attains the unique value $t = A^{-1}\xi$ with probability converging to 1, when T is extended to R^p. Taking into account the weak convergence of measures and continuity of the functional arg min (with its unique value for the random field in the limit), we get the asymptotic normality of $\sqrt{N}(\theta^V - \theta^*)$ with parameters $(0, A^{-1}BA^{-1})$.

Lemma 1 directly implies:

Proposition 1. $E[\sqrt{N}(\theta^V - \theta^*)]^u < C < \infty$ *for* $u \le \lceil r-2 \rceil$ *under assumptions of Section 3 and (A4.(r)). It converges to the moments of the limiting distribution for* $\sqrt{N}(\theta^V - \theta^*)$ *as* $N \to \infty$.

Let us emphasize that finding global minimum of (7) is in general a complicated computational task, i.e. construction of effective consistent estimator of θ^* remains an open problem.

5. REAGAN

LSE (7) for a Q-model with $D^{-1}(x,\theta)$ used instead of $V(x)$ is **generally not a consistent estimator** contrary to claims in some applied papers.

Theorem (**1**.1.b) implies that it converges in probability to

$$\arg\min_{\Theta}\{\int tr[D^*(x)D^{-1}(x,\theta)]d\varepsilon(x) + R^*_{D^{-1}(x,\theta)}(\theta)\}. \tag{10}$$

Let us give a definition of *locally asymptotically unbiased* procedure **REAGAN**.

Definition 1. Let $\theta^s = \theta^s(y) \in \Theta$ be an estimator at the sth step of the algorithm and observations $y = y^{(N)}$. Let us introduce the operator $(\mathbf{R}^p \to \mathbf{R}^p)$

$$A_N(\theta) \doteq (Nm_N(\theta))^{-1}\varphi^T D^{-1}(x_1^N, \theta),$$

$$\delta(\theta) = Vec\{\delta_1^N(\theta)\}, \qquad \varphi(\theta) = Vec\{\varphi(x_1^N, \theta)\},$$

and introduce

$$\theta^{s+1} = \mathfrak{A}_y(\theta^s) \doteq \theta^s + A_N(\theta^s)\delta(\theta^s). \tag{11}$$

It is easy to check that operator $\mathfrak{A}_y(\theta)$ and its derivatives $L_i = \partial\mathfrak{A}_y(\theta)/\partial\theta_i, \quad i = 1, \dots, p$ are well defined in probability under assumptions of section 3 without assumption (A3.c). An analogue of *stochastic continuity* holds for random variable $\tau_y(r) = \sup_{1\le i\le p}\sup_{B_r}\|L_i\|$.

Lemma 2. *For $\forall\ \kappa > 0$*

$$\lim_{r\to 0}\lim_{N\to\infty} P(\tau_y(r) > \kappa) = 0$$

Proof. Using easily verifiable identities $\partial\delta(\theta)/\partial\theta = -\varphi(\theta)$ and

$$A_N(\theta)\varphi(\theta) = \mathbf{I}, \tag{12}$$

we obtain:

$$\begin{aligned} L_i = I + (\partial A_N^{-1}/\partial\theta_i)\delta(\theta) + A_N(\theta)(\partial\delta(\theta)/\partial\theta_i) = \\ = (\partial A_N(\theta)/\partial\theta_i)(\delta^* + \Delta\mu(\theta^*, \theta)). \end{aligned} \tag{13}$$

Further,

$$\partial m_N^{-1}(\theta)\partial\theta_i = -m_N^{-1}(\theta)(\partial m_N/\partial\theta_i)m_N^{-1}(\theta)$$

converges to a limit as $N \to \infty$ because of assumption (3) and Theorem (**1**.1.b). Thus, it is uniformly bounded and continuous in $N > N_0$. The same is true on $X \times \Theta$ for the functions φ, $\partial\varphi/\partial\theta$, V, $\partial V/\partial\theta$ which are parts of expression L_i. Thus, we can apply Theorem (**1**.2) to (13) and get the desired result. Let us emphasize that $\tau_y(r)$ is measurable in r which is insured by the fact that the sup in definition of τ could be taken over rational points because of continuity of $L_i(\theta)$ in θ.

The next Lemma uses the *independence* of random variables $y_1, \ldots, y_N$.

Lemma 3. *Let us denote $\rho_y(\theta) = \mathfrak{A}_y(\theta) - \theta$. Then $\sqrt{N}\rho^* \Rightarrow \mathbf{N}(0, [m^*]^{-1})$.*

Proof. The proof follows from equality $\rho^* = \sqrt{N}A_N^*\delta^*$ and the Central Limit Theorem.

Let us point out that assumption (3) of Theorem 1 is **not sufficient in our model for asymptotic normality** of $\sqrt{N}\rho^*$ and thus for asymptotic normality of $\hat{\theta}$, θ^s, $\ s \geq 1$. The next statement establishes REAGAN convergence under a consistent initial guess for θ^*.

Theorem 3. *There exists a random variable $\hat{\theta}(y)$ defined in probability s.t.:*

(a) events $\Gamma_r^N = \{y : \sup_{\theta^0\in B_r} \left\|\theta^s - \hat{\theta}(y)\right\| \to 0\}$ hold in probability as $r = r_N \to 0$;

(b) more precisely, event Γ_r^N occurs, if

$$\tau_y(r) + \frac{\|\rho^*\|}{r} < 1; \tag{14}$$

(c) the following bound is valid under condition (14):

$$\sup_{\theta^0\in B_r} \left\|\theta^s - \hat{\theta}(y)\right\| \leq r\frac{\tau_y^s(r)}{(1 - \tau_y(r))};$$

(d) $\sqrt{N}(\hat{\theta}(y) - \theta^*)$ is bounded in probability, i.e.

$$\lim_{K\to\infty}\lim_{N\to\infty} P(\left\|\hat{\theta}(y) - \theta^*\right\| > \frac{K}{\sqrt{N}}) = 0;$$

(e) $\hat{\theta}(y)$ satisfies equation

$$A_N(\hat{\theta})\delta_y(\hat{\theta}) = 0. \tag{15}$$

Proof. (b) By Lemmas 2 and 3 we have

$$\sup_{\theta\in B_r}\|\mathfrak{A}_y(\theta)-\theta^*\| \le \sup_{B_r}\left\|\mathfrak{A}_y(\theta)-\mathfrak{A}_y^*\right\| + \left\|\mathfrak{A}_y^*-\theta^*\right\| \le$$
$$\le r\tau_y(r)+\|\rho^*\| \le r$$

under condition (14), i.e. **Contraction Mapping** Ψ_y maps B_r into itself. Next we can use the famous bound from **Contraction Mapping Principle**

$$\sup_{\theta^0\in B_r}\left\|\theta^s-\theta^t\right\| \le r\frac{\tau_y^s(r)}{(1-\tau_y(r))}$$

for $\forall t > s \in \mathbf{N}$, which implies convergence of θ^s for all $\theta^0 \in B_r$ to the unique point $\hat\theta(y)$ under (14); $\hat\theta(y)$ is a measurable function of y because θ^s (11) is a random variable for any θ^0. Thus, we obtain (b) and by taking the limit as $t\to\infty$ we also get (c). Let us define

$$T_N^K=\{y:\|\delta^*\|\le K\sqrt{N},\ \left\|\sqrt{N}\rho^*\right\|\le K\}.$$

Then for $\forall\ \varepsilon>0$, there exists K s. t. $\overline{\lim}P(T_N^K)>1-\varepsilon$. Then for $y\in T_N^K$, condition (14) holds whenever $N>N_0,\ r,\ N_0$ are s. t. $KU(r)+V(r)\le\frac{1}{2}$ (by Lemma 2) and $2K<r\sqrt{N}$ for $N>N_0$. It is obvious that solution of these two inequalities is $r\in[\frac{2K}{\sqrt{N}},r_0]$, where $r_0(K)$ is the maximal solution of the first inequality (the end points of the above interval correspond to results (a) and (d) respectively). Equality (15) is obtained by taking the limit in (11) and using the fact that $A_N(\theta)\delta(\theta)=\rho(\theta)$.

Remark 3. (1) In many previous works on convergence in probability of Gauss-Newton-type algorithms ([5–7]),"the proof"was essentially based on verification of the condition (similarly to the erroneous section 4.3.8 in [47]):

$$\sup_{B_r}\underline{\lim}_{N\to\infty}P\{\|\partial\mathfrak{A}_y(\theta)/\partial\theta\|\le\alpha<1\}\to 1$$

with the help of the standard Law of Large Numbers (which is weaker than (14)). The above condition does not necessarily imply the required condition:

$$P\{y:\left\|\Delta\mathfrak{A}_y(\theta^s,\theta^{s-1})\right\|\le\alpha\left\|\theta^s-\theta^{s-1}\right\|\ \forall\ s\in\mathbf{N}\}\to 1.$$

Moreover, the Law of Large Numbers cannot be applied directly to

$$\partial\mathfrak{A}_y(\theta^s)/\partial\theta=A_N(\theta^s)(\delta^*+\Delta\mu(\theta^*,\theta)),$$

because random variable θ^s depends on δ^* for $s\ge1$. The last mistake is also typical in deriving bounds for remainder terms, when proving asymptotic normality in similar works, where Theorem (**1**.3) should be applied instead of LLN.

(2) From inequalities for functions $g_i\in Lip(\rho),\ i=1,\dots,n$, it follows (see Section 7):

$$\sup_{\Theta}\left|\sum_{i=1}^n g_i(\theta)\delta_i^*\right|^2\le A(\delta)\sum\|g_i\|_\rho^2,\quad E[A^2]<\infty,$$

which, similarly to the proof of Theorem **1**, implies a **stronger bound**

$$\tau_y(r)\le\lambda+\frac{A_\lambda(\delta)}{\sqrt{N}}+V(r),\quad E[A_\lambda^2]<\infty\ \forall\ \lambda>0,$$

that allows to prove convergence in probability of $\theta^s \to \hat{\theta}(y)$ as in Theorem **3**, whenever $\left\|\theta^0 - \theta^*\right\| < r$ for some $r > 0$ (without requiring that $\theta^0 \to \theta^*$ in probability). Moreover, B_r is invariant under action of $\mathfrak{A}_y$ even without condition (A1.c) which by the Schauder fixed point theorem implies convergence of iterations in probability to $\sqrt{N}$-consistent estimate $\hat{\theta}$ of θ^*.

Theorem 4. *Under conditions of Lemma 2:*
(1) if θ^0 is a consistent estimate of θ^*, then $\sqrt{N}(\hat{\theta} - \theta^*) \Rightarrow \mathbf{N}(0, m^{-1}(\theta^*))$.
(2) If $\sqrt{N}\left\|\theta^0 - \theta^*\right\|$ is bounded in probability, then $\sqrt{N}(\theta^s - \theta^*) \Rightarrow \mathbf{N}(0, [m^*]^{-1})$.
(3) Under additional requirements of (A3.c) and (A4.(r)) with $r > 5$, taking $\theta^0 = \theta^V$ for some $V > 0$ from (A3.c) implies that the two first moments of $\sqrt{N}(\theta^1 - \theta^*)$ are bounded and converge to the corresponding moments of the limiting Normal distribution.

Proof. From (15) and (1), it is easy to obtain the following equality:

$$\sqrt{N}(\hat{\theta} - \theta^*) = \sqrt{N}\rho^* + R_y, \qquad R_y = \mathbf{I} + \mathbf{II},$$

$$\mathbf{I} \doteq \sqrt{N}\int_0^1 [\partial(A_N(\theta^* + \lambda(\hat{\theta} - \theta^*))\delta^*)/\partial\theta](\hat{\theta} - \theta^*)d\lambda,$$

$$\mathbf{II} \doteq \sqrt{N}A_N(\hat{\theta})\int_0^1 \Delta\varphi(\theta^* + \lambda(\hat{\theta} - \theta^*), \hat{\theta})(\hat{\theta} - \theta^*)d\lambda.$$

It is sufficient by Lemma 3 to check that $R_y \to 0$ in probability which follows immediately from Theorem (**3**.d) and Theorem (**1**.3).

To prove (2), we use (1) again to obtain the set of equalities

$$\mathfrak{A}_y(\theta^s) - \theta^* = \rho^* + R_y, \qquad \sqrt{N}R_y = \mathbf{I}^{(s)} + \mathbf{II}^{(s)}, \tag{16}$$

$$\mathbf{I}^{(s)} \doteq \sqrt{N}\int_0^1 [\partial(A_N(\theta^s + \lambda(\theta^* - \theta^s))\delta^*)/\partial\theta](\theta^s - \theta^*)d\lambda,$$

$$\mathbf{II}^{(s)} \doteq \sqrt{N}A_N(\theta^s)\int_0^1 \Delta\varphi(\theta^s + \lambda(\theta^* - \theta^s), \theta^s)(\theta^s - \theta^*)d\lambda.$$

Then we proceed with the proof by induction in s, taking into consideration boundedness in probability of $\sqrt{N}(\theta^s - \theta^*)$, in the same way as in Theorem (**4**.3). From (16) with $s = 0$ using Minkovski inequality for $r = 4$

$$E[\left|\sum \xi_i\right|^r] \leq (\sum (E[|\xi_i|^r])^{1/r})^r$$

and using the fact that $\sqrt{N}\left\|\theta^0 - \theta^*\right\|$ has bounded moments (from (1)), and finally applying Cauchy–Schwarz inequality, we can prove that $E[\left\|\mathbf{I}^{(1)}\right\|^2] < \infty,\ for\ N > N_0$. The boundedness for $N > N_0$ of the second moments of $\mathbf{II}^{(1)}$ and $\rho^*\sqrt{N}$ is proven as in Lemma 2 using the Whittle's inequality in the second case. The convergence of second moments follows immediately from Theorem (**1**.1.b) which in turn implies (3).

6. LOWER BOUND FOR LINEAR UPDATES AND LAM OF REAGAN

Here we establish Local Asymptotic Minimaxity (LAM) of REAGAN with respect to the class of linear in y updates. Let us fix a Q-model and a sequence of discrete designs ε_N satisfying assumptions of section 3 without (A1.c) and (A3.c). Let us define risk to be:

$$r_N^l(t, \theta^*) \doteq N \sup_{|l^T(\theta - \theta^*)| < C/\sqrt{N}} E_\theta[(l^T\theta - a^T y)^2].$$

Theorem 5. *The following lower bound for the square loss of linear in y estimates $t = a^T y$, $y = Vec\{y_1^N\}$ of the parameter $[\theta^*]^T l$, $l \in R^p$ is valid:*

$$r(t) \doteq \underline{\lim}_{C\to\infty}\underline{\lim}_{N\to\infty} r_N^l(t,\theta^*) \geq l^T m^{-1}(\theta^*) l. \tag{17}$$

Instead of $\underline{\lim}_{C\to\infty}$ in (17) we could use condition $C^2 > l^T m^{-1}(\theta^*) l$.

Proof. We have

$$E[(l^T\theta - t)^2] = (E_\theta[t] - l^T\theta)^2 + a^T D_\theta a.$$

Further,

$$a^T\mu(\theta) - l^T\theta = (a^T\mu^* - l^T\theta^*) + (a^T\varphi^* - l^T)(\theta-\theta^*) + \\ + a^T[\int \varphi(\theta^* + \lambda(\theta-\theta^*))d\lambda - \varphi^*](\theta-\theta^*) \doteq \mathbf{I} + \mathbf{II}(\theta-\theta^*) + \mathbf{III}.$$

If $\mathbf{I} \neq 0$ or $\mathbf{II} \neq 0$, then their contribution to $r_N^l(t,\theta^*)$ is of order N and $O(C)$ for $C \to \infty$ respectively. At the same time, the contribution of **III** is of order $O(C^2/N)$. Under condition of "local unbiasedness" of estimate t: $\mathbf{I} = \mathbf{II} = 0$, using method of proof of Gauss–Markov Theorem for linearized model

$$E_\theta[y - \mu^*] = \varphi^*(\theta - \theta^*), \quad Cov_\theta[y_i] = D_i^*$$

and Theorem (**1**.1.b), we obtain inequality

$$D_\theta[t] \geq N^{-1} l^T [m^*]^{-1} l(1 + o(1)),$$

which is an asymptotic equality (Gauss). This completes the proof of the theorem.

If additionally (A1.c) holds, then estimates $\hat\theta$, θ^s, $s \geq 1$ achieve their asymptotic limit (17) under conditions (**4**.1,2) in terms of characteristics of limiting distribution. Under condition (**4**.3), the same is true for θ^1 in terms of its moments also.

Estimate θ^1 is linear in y up to the fact that coefficients of correspondent linear form are determined by $\sqrt{N}$-consistent estimate θ^0. The LAM property of θ^1 means that for any $\sqrt{N}$-consistent initial approximation θ^1 is the most precise asymptotically among all linear in y estimates (even those depending on θ^*).

7. COROLLARIES OF THE TYPE 2 EMBEDDINGS OF BANACH SPACES

Let a continuous mapping $\rho : [0 : \infty) \to [0, \infty), \rho(0) = 0$ be s.t.

1. $\rho()$ determines a metric in C (satisfying the triangle inequality);

2. The minimal number $N()$ of ρ-balls of radius ε which cover $B(r), r > 0$, satisfies the Dudley–Strassen inequality

$$\int_0^a (\log N)^{1/2} d\varepsilon < \infty$$

for some $a > 0$.

Let Lip_ρ be the space of continuous functions on $B(r)$ with norm

$$||g||_\rho = g(\theta^*) + \sup_{\theta,\theta' \in B(r)} |g(\theta) - g(\theta')|/\rho(||\theta - \theta'||).$$

It is proved in [50] that the embedding $Lip(\rho) \to C(B(r))$ has type 2. This means that for every zero mean independent RV $X_1, \ldots, X_N$, the following inequality holds:

$$E||\sum X_i||^2_{B(r)} \leq A \sum E||X_i||_\rho$$

with the same constant for all $N, X_1, \dots, X_N$.

Let additionally to M1–M3 conditions hold: maximal continuity moduli of $\partial \mathfrak{A}_y(\theta)/$ satisfy inequality

$$\omega(\cdot) \le c\rho(\cdot) \tag{18}$$

for some $c > 0$.

Then the inequality (18) implies inequality

$$E||\partial A_N(\theta)/|| \le c/N$$

for some $c > 0$ and

$$\tau_y(r) \le \varepsilon/\sqrt{N} + V(r).$$

Remark. Condition (18) holds for $\rho(x) = x^b, 0 < b < 1$.

Theorem 6. *1. If conditions of theorem 3 (section 5) hold and initial approximation θ^0 is sufficiently close to θ^*, then the limit θ^∞ exists in probability and satisfies*

$$\sqrt{N}(\theta^\infty - \theta^0) \to N(0, M^{-1})(\theta^*);$$

2. The same convergence holds for $\theta^s, s \ge 1$, if $\sqrt{N}(\theta^0 - \theta^*)$ is bounded in probability;

3. The condition of 2. holds for $\theta^0 = \theta^W$, where the constant weight matrix W is positively definite. In this case, in addition to the statements of 2., two first moments of $\sqrt{N}(\theta^1 - \theta^*)$ are bounded in probability and converge to the same moments of the limiting distribution.

Proof. 1. The representation

$$\sqrt{N}(\theta^\infty - \theta^0) = \sqrt{N}\rho^* + R_y, R_y = \mathbf{I} + \mathbf{II},$$

$$\mathbf{I} = \sqrt{N}\int_0^1 [(\partial(A_N)/\partial\theta)(\theta^\infty + \lambda(\theta^\infty - \theta^*))\delta^*)]d\lambda,$$

$$\mathbf{II} = \sqrt{N}A_N(\theta^\infty)\int_0^1 \Delta\varphi d\lambda(\theta^* + \lambda(\theta^\infty - \theta^*), \theta^\infty)(\theta^\infty - \theta*),$$

follows from lemma 3 and theorem 3.3 (section 5). It is straightforward to prove that $R_y \to 0$ in probability using lemma 3 and theorem 3 (section 5).

Proof of 2: The Identity **1** and (11) imply

$$(\theta^s) - \theta^* = \rho^* + R_y,$$

$$R_y = I^{(s)} + II^{(s)}.$$

The proof uses induction over s and boundedness of $\sqrt{N}(\theta^1 - \theta^*)$ in probability and is similar to the proof of Theorem 3 (section 5). For $s = 0$ and for all sufficiently large N applying (19) and the Minkowski inequality

$$E|\sum \xi_i|^r \le (\sum (E|\xi_i|^r))^{1/r})^r$$

for $r = 4$ and statement of 1, we get $E^*||l^{(1)}||^2 < \infty$ and boundedness of $I^{(s)}$ and of second moments of $\sqrt{N}\rho^*$ using Theorem 3 (section 5) and the Whittle inequality.

The convergence of the second moments follows from Theorem 4 (section 5). and Theorem 5.2.1.b implying the statement of 3.

Comparing with Theorem 3 (section 5), we see that the lower bounds for θ^1 under static design ε_N become equalities under our conditions.

8. REAGAN PROPERTIES UNDER ADAPTIVE ASYMPTOTICALLY DETERMINISTIC DESIGNS

REAGAN procedures for adaptive asymptotically deterministic (AD) designs satisfying for some $r > 0$ and every $g(\cdot) \in (C(X) \times \Theta)$, the condition

$$\lim_{t\to\infty} (N)^{-2} D^* \sum_{i=1}^{N} g(x_i, \theta) = 0$$

uniformly over $\theta \in B(r)$ are studied in [28]. This condition is usually valid for *locally optimal sequential designs* for estimation or testing hypotheses.

Asymptotic Normality and LAM property of estimates is proved under this assumption in [28].

Remark 4. If this condition is violated, then instead of the Normal distribution of the limiting statistic, the **mixture of Normals** usually holds, see [29, 35].

9. LAM-POLYNOMIAL ESTIMATES. VARIANCE COMPONENTS.

Let us consider the following generalization of F-model—"F^k-model"omitting technical details similar to the ones above. Let the set $M_i(2k) = (m_i(1), \ldots, m_i(2k))$ of the first $2k$ moments of the observation y_i of dimension m is a function of $\theta \in \Theta \subset \mathbf{E}^p$ and $x_i \in X,\ i = 1, \ldots, N$, with the global (as in (A3.c)) or local (as in (A3.b)) requirement that θ is uniquely determined by $M_1(k), \ldots, M_N(k)$, and regularity condition similar to ones of sections 3 and 4. Then it is natural to study the class $\mathfrak{P}_k$ of estimators of θ^* of the form $t = \sum_i P_{ki}(y_i)$ where P_{ki} are polynomials of degree not higher than k taking estimator's second moment as optimality criterium (as in section 6). Let us introduce the vector

$$z_i = Vec\{y_i, Vec(y_i y_i^T), \ldots, Vec(y_i^{\otimes k})\}$$

where $a^{\otimes k} = a_1 \otimes \ldots \otimes a_k$ is a tensor product of $a_1 = \ldots = a_k = a$. It is clear that $E_\theta[z_i]$ is a function of $M_i(k)$, while $Cov_\theta[z_i]$ is a function of $M_i(2k)$. This is why measurements z_i are described by *extended* Q^2-model whose dimension A_{mk} does not exceed the number of different monomials in y_1 of degree not higher than k. Class $\mathfrak{P}_k$ is the same as the class of linear in $z_1, \ldots, z_N$ estimators of the form $t = l^T z$. This fact and regularity conditions similar to those in Theorem **4** imply the following result.

Theorem 7. REAGAN-*estimates for extended Q^2-model are LAM with respect to the class $\mathfrak{P}_k$ of estimates for Q^k-model.*

Let us explain how to reduce Variance Component Model to Q^2-model by using simple example. Consider

$$y_i = \varphi_i \beta + e_i + \alpha_i \mathbf{1}_{J_i},\ y_i, e_i \in \mathbf{R}^{J_i},\ \beta \in \mathbf{E}^p,\ \varphi_i \in \mathbf{R}^{J_i \times p}$$

$$J_i \leq J,\ \alpha_i \in \mathbf{E}^1,\ e_i = (e_{i1}, \ldots, e_{iJ_i})^T$$

where $\alpha_i,\ e_{ij}$ are mutually independent random variables.

$$E[e_{ij}] \equiv E[\alpha_i] \equiv 0,\ D[e_{ij}] \equiv \sigma^2,\ D[\alpha_i] = \sigma_A^2,\ E[y_i] = \varphi_i\beta,\ Cov(y_{ij}, y_{i'k}) \equiv 0\ \forall\ i \neq i',$$

$$Cov[y_i] = \sigma^2\mathbf{I} + \sigma_A^2\mathbf{1}_{J_i}\mathbf{1}_{J_i}^T,\ E[y_{ij}^k] = f_{ki}(\beta, \sigma, \sigma_A),\ k = 3, 4,$$

where $f_{ki},\ \varphi_i,\ \partial\varphi_i/\partial\theta,\ \partial f_{ki}/\partial\theta,\ \theta = (\beta, \sigma, \sigma_A)$ are continuous bounded functions. Then $y_i,\ i = 1, \ldots, J$ are described by F^2-model in each block $B_g = \{i : J_i = g\}$. Other **VCM**s are reduced to Q^2-models in a similar way. We display it in detail in our section 12 for proving profound results on one-way mixed ANOVA model.

10. REAGAN RELATION TO MLE

Q-models specify only two first moments of a distribution. The equation (15) is generally not equivalent to the Maximum Likelihood Equations, even for measurements distributed according to Gaussian (Normal) Distribution. It is worth considering the following problem:

Problem 1. For a given Q-model determined by $\mu(\theta)$ and relation between $V_i(\theta)$ and $\mu(\theta)$ find the distribution family $P_{x,\theta}$ such that estimators $\hat{\theta}$ for this distribution satisfy Maximum Likelihood equation or, more generally have the worst possible covariance matrix.

This family will then be asymptotically the worst in the sense of quadratic risk (under suitable regularity requirements) for the respective Q-model, but **REAGAN**-estimates will be asymptotically *MiniMax* estimates in the **class of arbitrary estimators**.

This is done so far in situations when relationships $\mu_i = \mu(x_i, \theta),\ D_i = D(x_i, \theta)$ represent 'curves' in **regular exponential family** of distributions $P_{x,\theta}^*$ with the density w.r.t. measure μ on $\mathbf{R}^m$:

$$p_{x,\theta}^*(y) \doteq \exp\{h^T(x, \theta)y - \nu(x, \theta)\},\ y \in \mathbf{R}^m \tag{19}$$

It is easy to check that in this case the Maximum Likelihood equation is reduced to the form:

$$\sum_{i=1}^N \frac{\partial \ln p_{x_i,\theta}^*(y_i)}{\partial\theta} = \sum_{i=1}^N \varphi^T(x_i, \theta)D^{-1}(x_i, \theta)\delta_i(\theta) = 0 \tag{20}$$

which is equivalent to (15). Under regularity requirements [2], the Maximum Likelihood equation has a unique solution which belongs to the domain of convergence of the integral of density (19). By Theorem 5.3 the solution $\hat{\theta}$ of the Maximum Likelihood equation (20) is the limit of **REAGAN** and if initial approximation θ^0 is $\sqrt{N}$-consistent then even θ^1 has the same asymptotic distribution as $\hat{\theta}$.

Let us consider class $\mathfrak{K}(\mu, D)$ of distributions $\tilde{P}_{x,\theta}$ with the same mean value as for (19) and covariance matrix $\tilde{D}(x, \theta) \leq D(x, \theta)$. Then under certain regularity conditions on densities $\tilde{p}_{x,\theta}$, MLE for them is also LAM estimate (in LAM definition we can take sup over domain $\|\theta - \theta^*\| \leq \frac{C}{\sqrt{N}}$ for sufficiently large C). At the same time **REAGAN**-estimate θ^1 for $\tilde{P} \in \mathfrak{K}(\mu, D)$ has the limiting quadratic risk which is no larger than one for density (19). Having in mind either quadratic risk for the limiting distribution or taking class $\mathfrak{K}(\mu, D)$ such that θ^1 has two finite moments (Theorem (**4**.3)) and with notation $r(t, \tilde{P})$ for the left side of (17) for arbitrary estimate t and distribution $\tilde{P}$, let us write down the following chain of inequalities:

$$\inf_t \sup_{\mathfrak{K}(\mu,D)} r(t, \tilde{P}) \geq \sup_{\mathfrak{K}(\mu,D)} \inf_t r(t, \tilde{P}) \geq$$

$$\geq \inf_t r(t, P^*) = r(\theta^1, P^*) \geq \sup_{\mathfrak{K}(\mu,D)} r(\theta^1, \tilde{P})$$

Thus, θ^1 is *MiniMax* LAM-estimator with respect to any estimator and distributions in class $\mathfrak{K}(\mu, D)$.

11. MULTIVARIATE NORMAL MEAN FIT UNDER UNKNOWN CONSTANT COVARIANCE

The multivariate Normal Distribution with unknown mean $\mu(x_i, \theta)$ and **constant** covariance matrix $D = W^{-1}$ is a regular exponential family with density

$$p^*_{x,\theta,W}(y) = [\det W]^{\frac{1}{2}}(2\pi)^{-\frac{m}{2}} \exp\{tr(\frac{A(\theta)W}{2})\},$$

$$A(\theta) = \sum_{i=1}^{N} \delta_i(\theta)\delta_i^T(\theta).$$

Introduce

$$\mathfrak{K} \doteq \{\tilde{P}_{x,\theta,W} : \tilde{E}[z] = E^*[z], \widetilde{Cov}[z] \leq Cov^*[z]\},$$

where

$$z \doteq Vec(y, Vec[yy^T])$$

is the class of regular densities. Let us apply previous results to this class. The following is the system of Maximum Likelihood equations for P^*:

$$tr[(\widehat{W}^{-1} - A(\hat{\theta}))\partial\widehat{W}] = 0, \tag{21}$$

$$tr[\partial A(\hat{\theta})\widehat{W}] = 0. \tag{22}$$

It follows from (21) that $\widehat{W} = A^{-1}(\hat{\theta})$, and (22) implies that $\det A(\hat{\theta}) = \min$, i.e. we obtained the equation for **REAGAN**'s limit. This estimate is *MiniMax* in the class of arbitrary estimates and $\tilde{P} \in \mathfrak{K}$ which follows from the previous theory.

12. MIXED GAUSSIAN ANOVA

12.1. Introduction and Outline of Main Results

Consider a classical mixed ANOVA model

$$\mathbf{y} = X\gamma + \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{s_i U_i} \phi_\mathbf{i}. \tag{23}$$

Here $\mathbf{y} = \mathbf{y}^N$ is $(N \times 1)$ - vector of measurements, $X = X_N$, $U_i = (U_i)_N$ are respectively $(N \times p)$ and $(N \times n_i)$ -matrices of known parameters, $\phi_\mathbf{i}$ is an $(n_i \times 1)$ -normally distributed random vector $\phi_\mathbf{i} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id_{n_i}})$, $;\gamma \in \mathbf{R^p}$ and $s_i \in \mathbf{R}_+$, $i = 1, \ \dots \ , k$, are unknown parameters, $\mathbf{Id_{n_i}}$ is identity matrix of dimension n_i.

It is clear that

$$E\mathbf{y} = X\gamma,$$

$$V = Cov\mathbf{y} \ = \ \sum_{i=1}^{k} \ s_i^2 \ G_i, \quad G_i = U_i U_i^T,$$

thus $d_i = s_i^2$ are called variance components.

The pioneer work, where ANOVA methods were applied to testing hypotheses on variance components for a balanced mixed model was [9], later R.Fisher devoted some attention to those models in his famous book [10]. Important contributions to this theory were made later by F. Yates, A.

Wald, C. Eisenhart, H. Sheffe, S.R. Searle, C.R. Rao, T.W. Anderson among many others. A critical ANOVA overview by A.N. Kolmogorov is reproduced in [38].

We develop maximum likelihood estimates (MLE) for the parameter $\theta = (\gamma^T, d_1, \dots, d_k)$ of the distribution $P_\theta^n(\mathbf{y})$ in model (1). We do not touch unbiased quadratic estimates studied in [41].

For mixed models MLE *cannot be evaluated analytically.* Asymptotics for MLE-s under some restrictive conditions was heuristically discussed in [13] Asymptotic normality (AN) of MLE-s under more general conditions was proved in [36].

The asymptotic efficiency and normality of MLE's *do not follow from Cramer's classical theory* because the *measurements in (1) are essentially dependent*

One of our principal aims in the present paper is to outline (for the simplest case of One-Way classification and bounded loss functions) a proof of the *Local Asymptotic Minimaxity* (LAM) of MLE and of certain approximations to MLE. LAM as formally defined further by the property just below (26), means that the deviation of the estimate from the true value θ^* is as minimal as possible in the local minimax sense.

We prove this here for One-Way ANOVA using certain *modified Fisher score updates* introduced further in section 12.2. Although the derivation **scheme** is similar to that displayed in sections 3–8, the *complexity* of this iterative **quadratic** estimation procedure grows considerably. Accordingly, the *enumeration of subsections and formulas is separate starting in this large section 12.*

For the convergence of our modified iterative Fisher score statistic in probability we need *arbitrary (non- qualified) consistency* of the initial guess. A geometric rate of the iterations' convergence uniformly over initial guesses from a neighborhood of θ^* proved by us in theorem 8 implies that after *const* $\log N$ iterations (for appropriately large *const*) we get the qualified consistency of the derived estimate. We show that *the next iteration provides us with an efficient (maximin) estimate* and that the *limit of iterations satisfies MLE equation* section 12.4. Thus, LAM property of MLE becomes transparent since the modified Fisher score update for MLE is MLE itself and the proved contraction property of the modified Fisher score iterations imply the convergence of any solution to the MLE equation in probability to a single point as the sample size increases.

The Local Asymptotic Normality (or simply LAN) [20] is (in our case) the following decomposition of

$$L_n(\mathbf{u}) = \ln[dP^{(n)}_{\theta+\Xi^{1/2}\mathbf{u}}/\ d\ P^{(n)}_\theta], \mathbf{u} \in \mathbf{R}^{p+k}:$$

$$L_n(\mathbf{u}) = u^T\lambda - (1/2)\mathbf{u}^T J\mathbf{u} + \psi_n(\mathbf{u}), \tag{24}$$

where

$$\lambda \sim N(0, J), J = \begin{Bmatrix} B & 0 \\ 0 & C, \end{Bmatrix} \quad d_i > 0,\ i = 1, \dots, k,$$

and $\psi_n(\mathbf{u})$ converges in $P_\theta^{(n)}$-probability to zero.

This decomposition was claimed in [32] for general mixed ANOVA-models under the conditions close to those of [36].

Particularly,
$$\Xi^{-1} = diag(\nu_0(n), \dots, \nu_{p-1}(n), \nu(n)_p, \dots, \nu(n)_{p+k-1},);$$

$$\nu_i(n) \equiv \nu_0(n), 0 \le i \le p-1,$$

was assumed in [32] to provide the existence for $1 \le i \le j \le k$ of

$$C_{ij} = \lim_{n\to\infty}(1/2)tr(V^{-1}G_iV^{-1}G_j)\nu_{p-1+i}^{-1/2}(n)\nu_{p-1+j}^{-1/2}(n), \quad B = \lim_{n\to\infty}\nu_0^{-1}(n)X^TV^{-1}X. \tag{25}$$

Complete proofs in [32] were not published. The parameters of the limiting normal distribution for MLE estimates found in [36] coincide with those in (25). We prove the existence of these limits in section 12.4 under mild regularity conditions.

A function $w(\cdot) : \mathbf{R}^p \to \mathbf{R}^+$ is called bowl-shaped if $\{\mathbf{u}\| \ w(\mathbf{u}) \le a\}$ are closed bounded symmetric convex sets for any $a \ge 0$.

The fundamental Hajek's lower bound for the LAM-risk of any estimate T^n (defined by the left-hand side of (26)) for any bowl-shaped loss function $w(\cdot)$

$$\liminf_{n\to\infty}\{\sup_{\theta\in\Theta} E_\theta w\big(J^{-1/2}(\theta)\cdot \Xi^{-1/2}(T^n-\theta)\big)\} \ge \int w(\mathbf{u})(2\pi)^{-(k+p)/2}e^{-|\mathbf{u}|^2}d\mathbf{u}, \tag{26}$$

is implied by LAN (see e.g. [15]). T^n satisfies LAM if equality holds in (26).

Our analysis in section 12.5 implies the *uniform* convergence of $\psi_n(\mathbf{u})$ from (24) to zero in $P_\theta^{(n)}$-probability. Namely, for all $K > 0, a > 0$

$$\lim_{n\to\infty} P_\theta^{(n)}\left(\sup_{||\mathbf{u}||<\mathbf{K}} |\psi_n(\mathbf{u})| > \mathbf{a}\right) = 0$$

holds.

This could be used for proving LAM of Le Cam's updates uniformly over some region of qualified initial guesses [20].

We rely on the theorem in our subsection 1 proving that the result θ^1 of the one-step iteration for a general multivariate regression model introduced in section 12.2 (see (30)) attains the equality in (26) for any bounded (not necessarily bowl-shaped) loss function uniformly over some set of $\Xi^{1/2}$-consistent initial guesses for θ, namely,

$$\lim_{n\to\infty}\sup_{\theta\in\Theta} E_\theta w\big(J^{-1/2}(\theta)\cdot \Xi^{-1/2}(\theta^1-\theta)\big) = \int w(\mathbf{u})(2\pi)^{-(k+p)/2}e^{-|\mathbf{u}|^{\mathbf{2}}}d\mathbf{u}. \tag{27}$$

In view of the Hajek's lower bound (4) this means Local Asymptotic Minimaxity (LAM) of θ^1.

The uniformity over initial guesses mentioned above permits us to plug in an initial guess depending on the same sample as the improvement θ^1 does.

We formulate and prove LAN property for One-Way ANOVA model in subsection 12.5. Before that we define and study in sections 12.3–12.4 a quadratic modification of iterative procedure REAGAN for fitting One-Way ANOVA model considered as a **multivariate regression model**. REAGAN uses modified Fisher-score statistics as the updates for the preceding step of this iterative procedure. We prove the convergence of REAGAN-estimates in probability to MLE in section 12.4. In [21] this was written for a general mixed ANOVA-model following the general method developed in [28]. [21] has not been published and seems to contain some gaps. Our method differs in some details from that of [21]. Particularly, for One-Way layout we diagonalize globally the covariance matrix of the the observations in section 12.3 which simplifies our analysis considerably as compared to a general mixed model, where this diagonalization can be made only locally under the true values of the parameters.

Our computation of the limiting covariance matrix J for the estimates of parameters in One-Way mixed model implies that the **designs minimizing natural functionals of** J **differ drastically** from those for One-Way model *with fixed parameters.* Namely, the optimal design **need not be balanced** to obtain the minimal principal term of the expansion of the risk in section 12.6, see also [1].

12.2. Assumptions. Definition of Iterative Procedure.

The following One-Way Mixed model is considered

$$y_{i,j} = \theta_0 + \theta_1 \cdot x_i + b_i + \varepsilon_{i,j}, \tag{28}$$

where we have n blocks $(1 \le i \le n)$ and each of the blocks contains precisely m_i elements (i.e. for a fixed i, $1 \le j \le m_i$). The size of each block m_i and the total number of observations $N = \Sigma_{i=1}^{i=n} m_i$ are positive nondecreasing integer-valued functions of n. The following assumptions are made about the model:

Assumption 1.1. There exist a constant $a > 0$ and a nondecreasing function $M = M(n)$ such that $0 < a \cdot M(n) \le m_i(n)\ \forall i \le n$.

Assumption 1.2. $\varepsilon = (\varepsilon_{1,1}, \ \dots, \varepsilon_{1,m_1}, \varepsilon_{2,1}, \ \dots, \varepsilon_{n,m_n})^T \sim \mathcal{N}(0, \beta \cdot \mathbf{Id}_N)$, $\mathbf{b} = (b_1, \ \dots, b_n)^T \sim \mathcal{N}(0, \alpha \cdot \mathbf{Id_n})$, where $N = \Sigma_{i=1}^{i=n} m_i$; the elements of $\mathbf{b}$ and ε are mutually and jointly independent.

We can rewrite (28) in matrix form as

$$\mathbf{y} = \mathbf{F}(x, \ \theta_0, \ \theta_1) + \mathbf{U} \cdot \mathbf{b} + \varepsilon. \tag{29}$$

Here $\mathbf{y} = (y_1, \ \dots, y_N)^T = (y_{1,1}, \ \dots, y_{1,m_1}, y_{2,1}, \ \dots, y_{n,m_n})^T$, $\mathbf{F}(x, \ \theta_0, \ \theta_1)$ is a column whose first m_1 entrees all equal $\theta_0 + \theta_1 \cdot x_1$ followed by $m_2 \quad (\theta_0 + \theta_1 \cdot x_2)$ and so on, and $\mathbf{U}$ is an $N \times n$ matrix that consists of zeros and ones, the first column being filled with m_1 ones first, followed by zeros, the second– with m_1 zeros, then with m_2 ones followed by zeros, etc., and, finally, the very last column of $\mathbf{U}$ has ones as the last m_n entries and zeros everywhere else; $\varepsilon = (\varepsilon_1, \ \dots, \varepsilon_N)^T = (\varepsilon_{1,1}, \ \dots, \varepsilon_{1,m_1}, \varepsilon_{2,1}, \ \dots, \ \varepsilon_{n,m_n})^T$.

Assumption 1.3.$\forall i \ x_i \in X$, and X is a bounded subset of $\mathbf{R}$.

Define θ^* to be the unknown true vector value of parameter

$$\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*)^T = (\theta_0^*, \theta_1^*, \alpha^*, \beta^*)^T$$

Our objective is to estimate θ^* given a vector of observations $\mathbf{y}$. The next assumption refers to the parameter space $\Theta \subset \mathbf{R}^4$.

Assumption 1.4. Θ is compact, contains an open neighborhood of θ^*, and $\forall \theta = (\theta_0, \theta_1, \theta_2, \theta_3)^T \in \Theta$ we have $\inf_{\theta \in \Theta} \alpha, \inf_{\theta \in \Theta} \beta > 0$ $(\alpha = \theta_2, \beta = \theta_3)$.

We take the $N \times 1$ vector $\mathbf{y}$ and construct from it the $N + N(N+1)/2$ vector $\mathbf{z}$ by adding products $y_k y_l$, $1 \le k \le l \le N$ in the lexicographic order:

$$\mathbf{z} = (y_1, \ \dots, y_N, y_1 y_1, \ \dots, y_1 y_N, \ \dots, y_{N-1} y_N, y_N y_N)^T.$$

The position in a row or a column of length $N + N(N+1)/2$ having the same number as that containing $y_k y_l$ (or z_p) in $\mathbf{z}$ if we count from the top will be referred to as the position that corresponds to $y_k y_l$, $1 \le k \le l \le n$ (or z_p, $1 \le p \le N + N(N+1)/2$); the same convention for “the position corresponding to y_l, $1 \le l \le N$”.

The obvious formula $\mathbf{z} = E\mathbf{z} + \mathbf{r}, \mathbf{Er} = \mathbf{0}$, where the first N entries of $E\mathbf{z}$ coincide with those of $\mathbf{F}(x, \ \theta_0, \ \theta_1)$, whereas the remaining components of $E\mathbf{z}$ are identical to those of the vector (for the definition of vech, see e.g. p. 332 in [43]).

$$vech \left\{ \mathbf{F}(x, \ \theta_0, \ \theta_1) \mathbf{F}^T(x, \ \theta_0, \ \theta_1) + \theta_3 \mathbf{UU^T} + \theta_4 \mathbf{Id} \right\},$$

and therefore represent quadratic functions of $\theta_0^*, \ \theta_1^*$. The covariance function of unbiased “errors” $\mathbf{r}$ in the above multivariate regression model is evaluated via first 4 moments of the multivariate gaussian distribution given its first 2 moments.

The Fisher score update for the parameters' initial guess of our model is given by the one step of the REAGAN. We show at the end of section 12.4 that after sufficiently many iterations we get arbitrarily close approximation to MLE, i.e. we collect essentially all information on the parameters of the model from the observations.

The REAGAN consists of iterations of the same one-step operator which is defined below:

$$\mathcal{U}(\theta) = \mathcal{U}_{n,\ \mathbf{z}}(\theta) = \theta + [\phi^T \cdot D^{-1} \cdot \phi]^{-1} \cdot \phi^T \cdot D^{-1} \cdot (\mathbf{z} - E\mathbf{z}). \tag{30}$$

The factors in (30) are:

$$\phi = \phi_{n,\ \mathbf{z}}(\theta) = \left[\frac{\partial}{\partial\theta} E\mathbf{z}\right]^T,$$

$$D = D_{n,\ \mathbf{z}}(\theta) = Cov(\mathbf{z},\ \mathbf{z}).$$

The matrix $\phi^T D^{-1}\phi$ in (30) is given in explicit form by formula (42) (replace $\tilde{\alpha}$, $\tilde{\beta}$ with α, β), so our next assumption pertains to the way x_i's should be chosen:

Assumption 1.5. For all n large enough

$$\inf_{\Theta} \left| \begin{Bmatrix} (1/n) \cdot \mathbf{Id}_3 & \mathbf{0} \\ \mathbf{0} & 1/N \end{Bmatrix} \cdot \phi^T D^{-1}\phi \right| > c > 0, \text{ where } c \text{ is a constant.}$$

(notice that $|\phi^T D^{-1}\phi|$ is always nonnegative).

Finally, the asymptotic results of Section 12.5 require two more assumptions.

Assumption 1.6. The limits $\lim\limits_{n\to\infty} \dfrac{1}{n}\Sigma x_i$ and $\lim\limits_{n\to\infty} \dfrac{1}{n}\Sigma x_i^2$ exist, are finite, and equal $\bar{x}$ and $\bar{x^2}$ respectively.

Assumption 1.7. $\lim\limits_{n\to\infty} M(n) = +\infty$ (i.e. $\inf_{1\le i\le n} m_i(n) \to +\infty$ as $n \to \infty$).

Notice that Assumptions 1.1 and 1.7 entail $\lim\limits_{n\to\infty} n/N = 0$.

Remark 5. Suppose that Assumptions 1.1–1.4, 1.6–1.7 hold. Then, (42) implies that Assumption 1.5 translates into a simple requirement that $\bar{x^2} - \bar{x}^2 > 0$.

12.3. Preliminary results.

We start by observing that the covariance matrix of $\mathbf{y}$ given by

$$V = V_{n,\ \mathbf{y}}(\theta) = Cov_\theta(\mathbf{y}, \mathbf{y}) = \alpha \cdot \mathbf{U}\mathbf{U}^T + \beta \cdot \mathbf{Id_N} \tag{31}$$

has a block-diagonal form. There are n blocks of dimensions $m_1 \times m_1,\ \dots\ , m_n \times m_n$ respectively and each of the blocks has $\alpha + \beta$ along the main diagonal and α's everywhere else:

$$\begin{Bmatrix} \alpha+\beta & \alpha & \dots & \alpha \\ \alpha & \alpha+\beta & \dots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \dots & \alpha+\beta \end{Bmatrix}. \tag{32}$$

V is symmetric and positive-definite. It is clear from (31) that there exists a block-diagonal orthogonal map $\mathcal{O}$ with block sizes identical to those of V, such that $\mathcal{O}^T V \mathcal{O}$ is a diagonal matrix *for all values of* $\theta \in \Theta$. It is an exercise in linear algebra to verify that an $m_i \times m_i$ block of type (32) has two eigenvalues, i.e. β and $m_i \cdot \alpha + \beta$, that occur with multiplicities $m_i - 1$ and 1 respectively. Set $\mathbf{y}^\circ = (y_1^\circ, \ \dots \ , y_N^\circ)^T = \mathcal{O}^T \cdot \mathbf{y}$. Without loss of generality we can assume that $\mathcal{O}$ is an $N \times N$ block-diagonal orthogonal matrix s.t. for $1 \le k \le N$,

$$Var_\theta(y_k^\circ) = m_i \cdot \alpha + \beta, \tag{33}$$

if $k = m_1 + \cdots + m_i, \beta$ otherwise

Next we observe that for an arbitrary linear automorphism B of $\mathbf{R}^{N+N(N+1)/2}$ and any $(N + N(N+1)/2) \times 1$ vector $\mathbf{t}$ filled with some constants, the operator $\mathcal{U} = \mathcal{U}_{n, \ \mathbf{z}}(\theta)$ is invariant with respect to replacement of $\mathbf{z}$ by $B \cdot \mathbf{z} + \mathbf{t}$:

$$\mathcal{U}_{n, \ \mathbf{z}}(\theta) \equiv \mathcal{U}_{n, \ B\cdot\mathbf{z}+\mathbf{t}}(\theta) \tag{34}$$

in fact, we even have

$$\phi_{n, \ \mathbf{z}}^T(\theta) \cdot D_{n, \ \mathbf{z}}^{-1}(\theta) \cdot \phi_{n, \ \mathbf{z}}(\theta) \equiv \phi_{n, \ B\cdot\mathbf{z}+\mathbf{t}}^T(\theta) \cdot D_{n, \ B\cdot\mathbf{z}+\mathbf{t}}^{-1}(\theta) \cdot \phi_{n, \ B\cdot\mathbf{z}+\mathbf{t}}(\theta) \phi_{n, \ \mathbf{z}}^T(\theta) \cdot D_{n, \ \mathbf{z}}^{-1}(\theta) \equiv$$

$$\phi_{n, \ B\cdot\mathbf{z}+\mathbf{t}}^T(\theta) \cdot D_{n, \ B\cdot\mathbf{z}+\mathbf{t}}^{-1}(\theta) \cdot \mathbf{B}. \tag{35}$$

Now, the idea is to use a suitable transformation of $\mathbf{z}$ that leaves the operator $\mathcal{U}$ invariant, but at the same time makes it much easier to analyze $\mathcal{U}$. For each $\theta \in \Theta$ we produce a separate transformation of the kind $\mathbf{z} \mapsto B\cdot\mathbf{z}+\mathbf{t}$ and compute $\phi_{n, \ \mathbf{z}}^T(\theta) \cdot D_{n, \ \mathbf{z}}^{-1}(\theta) \cdot \phi_{n, \ \mathbf{z}}(\theta)$ and $\phi_{n, \ \mathbf{z}}^T(\theta) \cdot D_{n, \ \mathbf{z}}^{-1}(\theta)$ explicitly.

Fix a $\tilde{\theta} \in \Theta$. Set $\tilde{\mathbf{y}} = (\tilde{y}_1, \ \dots \ , \tilde{y}_N)^T = \mathcal{O}^T(\mathbf{y} - \tilde{\mathbf{F}})$, here $\tilde{\mathbf{F}} = \mathbf{F}(x, \tilde{\theta}_0, \tilde{\theta}_1)$. Then $\tilde{\mathbf{y}}$ is a Gaussian vector with diagonal Covariance matrix and $E_{\tilde{\theta}}\tilde{\mathbf{y}} = 0$. Now, for some nonsingular matrix $\tilde{B}$,

$$\tilde{\mathbf{z}} = (\tilde{y}_1, \ \dots \ , \tilde{y}_N, \tilde{y}_1\tilde{y}_1, \tilde{y}_1\tilde{y}_2, \ \dots \ , \tilde{y}_N\tilde{y}_N)^T = \tilde{B} \cdot \mathbf{z} + \tilde{\mathbf{t}}. \tag{36}$$

Let $\tilde{\mathbf{F}}^\circ = \mathcal{O}^T \cdot \tilde{\mathbf{F}}$, then the kth component $\tilde{f}_k^\circ$ of this $N \times 1$ vector is

$$\tilde{f}_k^\circ = \sqrt{m_i} \cdot (\tilde{\theta}_0 + \tilde{\theta}_1 \cdot x_i), \tag{37}$$

if $k = m_1 + \dots + m_i$, 0 otherwise.

This is implied by the fact that for each $m_i \times m_i$ block of type (32) the one-dimensional eigenspace corresponding to the value $m_i \cdot \alpha + \beta$ is generated by the sum of all basis vectors of the subspace on which the block acts. $\tilde{\mathbf{B}}$ maps y_k to y_k°, $1 \le k \le N$, and $y_k y_l$ to $y_k^\circ y_l^\circ - \tilde{f}_k^\circ \cdot y_l^\circ - \tilde{f}_l^\circ \cdot y_k^\circ$, $1 \le k \le l \le N$, hence it can be represented as a product of two matrices:

$$\tilde{\mathbf{B}} = \begin{Bmatrix} \mathbf{Id}_N & \mathbf{0} \\ \mathbf{M}_{\tilde{\mathbf{F}}^\circ} & \mathbf{Id}_{N+\frac{N(N+1)}{2}} \end{Bmatrix} \cdot \begin{Bmatrix} \mathcal{O}^T & \mathbf{0} \\ \mathbf{0} & \left[Sym^2\mathcal{O}\right]^T \end{Bmatrix}, \tag{38}$$

where $Sym^2\mathcal{O}$ is the restriction of $\mathcal{O} \otimes \mathcal{O}$ to the subspace of $E \otimes E$ ($E =< e_1, \ \dots \ , e_N >$ being a linear space on which $\mathcal{O}$ acts) spanned by all vectors of the form $\frac{1}{2}(e_k \otimes e_l + e_l \otimes e_k)$, $1 \le k, l \le N$, and

$$\mathbf{M}_{\tilde{\mathbf{F}}^\circ} \tag{39}$$

is an $(N + \frac{N(N+1)}{2}) \times N$ matrix obtained from $\mathbf{Id}_N \otimes \tilde{\mathbf{F}}^\circ + \tilde{\mathbf{F}}^\circ \otimes \mathbf{Id}_N$ by removing $(N+1)st, \ (2N+1)st, \ (2N+2)nd, \ \dots \ , (k \cdot N + 1)st, \ \dots \ , \ (k \cdot N + k)th, \ ((k+1) \cdot N + 1)st, \ \dots \ , ((N-1) \cdot N + 1)st, \ \dots \ , ((N-1) \cdot N + N - 1)st$ rows.

$$\tilde{\mathbf{B}}^{-1} = \left\{ \begin{matrix} \mathcal{O} & \mathbf{0} \\ \mathbf{0} & (\left[Sym^2\mathcal{O}\right]^T)^{-1} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \mathbf{Id}_N & \mathbf{0} \\ -\mathbf{M}_{\tilde{\mathbf{F}}^\circ} & \mathbf{Id}_{N+\frac{N(N+1)}{2}} \end{matrix} \right\}. \tag{40}$$

It follows from the definition of $\tilde{\mathbf{z}}$ that

$$D_{n,\ \tilde{B}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta}) = diag(Var_{\tilde{\theta}}(\tilde{y}_1),\ \dots\ ,Var_{\tilde{\theta}}(\tilde{y}_N), Var_{\tilde{\theta}}(\tilde{y}_1\tilde{y}_1), Var_{\tilde{\theta}}(\tilde{y}_1\tilde{y}_2),\ \dots\ ,Var_{\tilde{\theta}}(\tilde{y}_N\tilde{y}_N))$$

and with a little bit more work involved one can show that

$$\phi_{n,\ \tilde{\mathbf{B}}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta}) = \left\{ \begin{matrix} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{m_1} & \sqrt{m_1}x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \sqrt{m_n} & \sqrt{m_n}x_n & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & m_1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & m_n & 1 \end{matrix} \right\},$$

where the first two columns have m_i and $m_i x_i$, $1 \leq i \leq n$, respectively in positions that correspond to $\tilde{y}_{m_1+\dots+m_i}$ and zeros everywhere else; the nonzero elements of the third column are precisely m_i, their positions correspond to $\tilde{y}_{m_1+\dots+m_i}\tilde{y}_{m_1+\dots+m_i}$. Finally, the last column contains ones corresponding to $\tilde{y}_k\tilde{y}_k$, $1 \leq k \leq N$, and zeros in every other position.

For a random vector $\xi = (\xi_1,\ \dots\ ,\xi_n)^T \sim \mathcal{N}(0, V)$ the covariance matrix of the vector $\xi \otimes \xi$ is given by $2N_n \cdot (V \otimes V)$, where N_n is an $n^2 \times n^2$ matrix such that for any $n \times n$ matrix A the following identity holds $N_n \cdot vecA = vec\left\{\frac{1}{2}(A + A^T)\right\}$ (Theorem 10.2, p.164 in [22]). This fact, together with the diagonality of $D_{n,\ \tilde{B}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})$ makes it possible to obtain $\phi^T_{n,\ \tilde{\mathbf{B}}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta}) \cdot D^{-1}_{n,\ \tilde{B}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})$:

$$\left\{ \begin{matrix} 0 \dots & \frac{\sqrt{m_n}}{\tilde{\alpha}m_n+\tilde{\beta}} & 0 & \dots & 0 & \dots & 0 \\ 0 \dots & \frac{\sqrt{m_n}x_n}{\tilde{\alpha}m_n+\tilde{\beta}} & 0 & \dots & 0 & \dots & 0 \\ 0 \dots & 0 & 0 & \dots & \frac{m_1}{2(\tilde{\alpha}m_1+\tilde{\beta})^2} & \dots & \frac{m_n}{2(\tilde{\alpha}m_n+\tilde{\beta})^2} \\ 0 \dots & 0 & \frac{1}{2\tilde{\beta}^2} & \dots & \frac{1}{2(\tilde{\alpha}m_1+\tilde{\beta})^2} & \dots & \frac{1}{2(\tilde{\alpha}m_n+\tilde{\beta})^2} \end{matrix} \right\}, \tag{41}$$

here, just like in the case of (40), nonzero elements only occur in positions corresponding to $\tilde{y}_{m_1+\dots+m_i}$ for the first two rows, $\tilde{y}_{m_1+\dots+m_i}\tilde{y}_{m_1+\dots+m_i}$ and $\tilde{y}_k\tilde{y}_k$ for the third and fourth rows respectively ($1 \leq i \leq n$, $1 \leq k \leq N$). In the first three rows these positions are filled with $\frac{\sqrt{m_i}}{\tilde{\alpha}m_i+\tilde{\beta}}$, $\frac{\sqrt{m_i}x_i}{\tilde{\alpha}m_i+\tilde{\beta}}$, and $\frac{m_i}{2(\tilde{\alpha}m_i+\tilde{\beta})^2}$ accordingly. For the fourth row, if $k = m_1 + \ldots + m_i$, then the position corresponding to $\tilde{y}_k\tilde{y}_k$ is filled with $\frac{1}{2(\tilde{\alpha}m_1+\tilde{\beta})^2}$, otherwise it has $\frac{1}{2\tilde{\beta}^2}$.

Next we exhibit $\phi^T_{n,\ \tilde{\mathbf{B}}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})\cdot D^{-1}_{n,\ \tilde{B}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})\cdot\phi_{n,\ \tilde{\mathbf{B}}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})$:

$$\left\{\begin{matrix} \Sigma\frac{m_i}{\tilde{\alpha}m_i+\tilde{\beta}} & \Sigma\frac{m_ix_i}{\tilde{\alpha}m_i+\tilde{\beta}} & 0 & 0 \\ \Sigma\frac{m_ix_i}{\tilde{\alpha}m_i+\tilde{\beta}} & \Sigma\frac{m_ix_i^2}{\tilde{\alpha}m_i+\tilde{\beta}} & 0 & 0 \\ 0 & 0 & \Sigma\frac{m_i^2}{2(\tilde{\alpha}m_i+\tilde{\beta})^2} & \Sigma\frac{m_i}{2(\tilde{\alpha}m_i+\tilde{\beta})^2} \\ 0 & 0 & \Sigma\frac{m_i}{2(\tilde{\alpha}m_i+\tilde{\beta})^2} & \Sigma\frac{1}{2(\tilde{\alpha}m_i+\tilde{\beta})^2}+\frac{1}{2\tilde{\beta}^2}\cdot(N-n) \end{matrix}\right\}, \tag{42}$$

the summation runs over $1 \leq i \leq n$. As the final step of preparation for evaluating the $\mathcal{U}(\theta)$ explicitly, we note that

$$\phi^T_{n,\ \mathbf{z}}(\tilde{\theta})\cdot D^{-1}_{n,\ \mathbf{z}}(\tilde{\theta}) \equiv \phi^T_{n,\ \tilde{\mathbf{B}}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})\cdot D^{-1}_{n,\ \tilde{B}\cdot\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})\cdot\tilde{\mathbf{B}}\cdot\mathbf{B}^{*-1}\mathbf{B}^*,\ \mathbf{B}^* \text{ is } \tilde{\mathbf{B}} \text{ when } \tilde{\theta}=\theta^*.$$

We now put together (41), (38), (40), and (39) to obtain $\phi^T_{n,\ \mathbf{z}}(\tilde{\theta})\cdot D^{-1}_{n,\ \mathbf{z}}(\tilde{\theta})$:

$$\left\{\begin{matrix} 0\ldots & \frac{\sqrt{m_n}}{\tilde{\alpha}m_n+\tilde{\beta}} & 0 & \ldots & 0 & \ldots & 0 \\ 0\ldots & \frac{\sqrt{m_n}x_n}{\tilde{\alpha}m_n+\tilde{\beta}} & 0 & \ldots & 0 & \ldots & 0 \\ 0\ldots & \frac{m_n\sqrt{m_n}(\theta_0^*-\tilde{\theta}_0+(\theta_1^*-\tilde{\theta}_1)x_n)}{(\tilde{\alpha}m_n+\tilde{\beta})^2} & 0 & \ldots & \frac{m_1}{2(\tilde{\alpha}m_1+\tilde{\beta})^2} & \ldots & \frac{m_n}{2(\tilde{\alpha}m_n+\tilde{\beta})^2} \\ 0\ldots & \frac{\sqrt{m_n}(\theta_0^*-\tilde{\theta}_0+(\theta_1^*-\tilde{\theta}_1)x_n)}{(\tilde{\alpha}m_n+\tilde{\beta})^2} & \frac{1}{2\tilde{\beta}^2} & \ldots & \frac{1}{2(\tilde{\alpha}m_1+\tilde{\beta})^2} & \ldots & \frac{1}{2(\tilde{\alpha}m_n+\tilde{\beta})^2} \end{matrix}\right\}\cdot\mathbf{B}^*, \tag{43}$$

the left multiplier in (43) differs from the matrix (41) only in the third and fourth rows: in (43), the third and the fourth rows contain nonzero elements in positions that correspond to $\tilde{y}_{m_1+\ldots+m_i}$, and those elements are precisely $\frac{m_i\sqrt{m_i}(\theta_0^*-\tilde{\theta}_0+(\theta_1^*-\tilde{\theta}_1)x_i)}{(\tilde{\alpha}m_i+\tilde{\beta})^2}$ and $\frac{\sqrt{m_i}(\theta_0^*-\tilde{\theta}_0+(\theta_1^*-\tilde{\theta}_1)x_i)}{(\tilde{\alpha}m_i+\tilde{\beta})^2}$ respectively, whereas these places are filled with zeros in the case of (41); all the other entrees of (41) and the left multiple of (43) are identical. Here, as usually, $1 \leq i \leq n$. It is convenient to introduce special notation for the left multiple in (43), hereafter it will be denoted by $\mathbf{Q}=\mathbf{Q}_{n,\ \mathbf{z}}(\tilde{\theta})$:

$$\phi^T_{n,\ \mathbf{z}}(\tilde{\theta})\cdot D^{-1}_{n,\ \mathbf{z}}(\tilde{\theta}) = \mathbf{Q}\cdot\mathbf{B}^*. \tag{44}$$

We are now ready to rewrite $\mathcal{U}$ defined by (12.1.3) in a form that simplifies our analysis in the subsequent section: for any $\theta\in\Theta$,

$$\mathcal{U}(\theta) = \theta + [\phi^TD^{-1}\phi]^{-1}\cdot\mathbf{QB}^*((\mathbf{z}-E_{\theta^*}\mathbf{z})+(E_{\theta^*}\mathbf{z}-E_\theta\mathbf{z})) =$$

$$\theta + [\phi^TD^{-1}\phi]^{-1}\cdot\mathbf{Q}\cdot(\mathbf{z}^*-E_{\theta^*}\mathbf{z}^*)+[\phi^TD^{-1}\phi]^{-1}\cdot\mathbf{QB}^*(E_{\theta^*}\mathbf{z}-E_\theta\mathbf{z}), \tag{45}$$

where, in view of (35), $[\phi^TD^{-1}\phi]$ is given by (42), $\mathbf{Q}$ is the left multiple in (43) (we replace $\tilde{\theta}$ with θ in $\mathbf{Q}$), and $\mathbf{B}^*$ is obtained from (38) by plugging in θ^* instead of $\tilde{\theta}$; $\mathbf{z}^*$ is obtained from $\tilde{\mathbf{z}}$ likewise. As the last result of this section, we compute the 4×1 vector $\mathbf{QB}^*(E_{\theta^*}\mathbf{z}-E_\theta\mathbf{z}) = \phi^T_{N,\ \mathbf{z}}(\theta)\cdot D^{-1}_{N,\ \mathbf{z}}(\theta)(E_{\theta^*}\mathbf{z}-E_\theta\mathbf{z})$:

$$\left\{\begin{matrix} \Sigma\frac{m_i(\theta_0^*-\theta_0+(\theta_1^*-\theta_1)x_i)}{\alpha m_i+\beta} \\ \Sigma\frac{m_ix_i(\theta_0^*-\theta_0+(\theta_1^*-\theta_1)x_i)}{\alpha m_i+\beta} \\ \Sigma\frac{m_i^2(\theta_0^*-\theta_0+(\theta_1^*\theta_1)x_i)^2+m_i^2(\alpha^*-\alpha)+m_i(\beta^*-\beta)}{2(\alpha m_i+\beta)^2} \\ \Sigma\frac{m_i(\theta_0^*-\theta_0+(\theta_1^*-\theta_1)x_i)^2+m_i(\alpha^*-\alpha)+(\beta^*-\beta)}{2(\alpha m_i+\beta)^2}+\frac{\beta^*-\beta}{2\beta^2}(N-n) \end{matrix}\right\}, \tag{46}$$

the sums are taken over $1 \leq i \leq n$.

12.4. Convergence of REAGAN to MLE.

In this section, using formulas obtained in 12.3, we establish the convergence of the REAGAN, provided that our initial approximation of θ^* belongs to some open neighborhood $\mathcal{B}^* \subset \Theta$ of θ^*. Unfortunately, we can tell nothing more about $\mathcal{B}^*$ than it exists. The proof depends heavily on

Theorem 8. *Under Assumptions 1.1–1.5, for any $r > 0$ there exists an open neighborhood $\mathcal{B}^*_r \subset \Theta$ of θ^* such that*

$$\lim_{n\to\infty} \mathbf{P}^n_{\theta^*} \left[\sup_{\mathcal{B}^*_r} \| \frac{\partial}{\partial\theta_j} \mathcal{U}(\theta) \| > r \right] = 0, \ \textit{where}\ 0 \le j \le 3. \tag{47}$$

In view of (45)

$$\frac{\partial}{\partial\theta_j}\mathcal{U}(\theta) = \frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \cdot \mathbf{Q} \right\} \cdot (\mathbf{z}^* - E_{\theta^*}\mathbf{z}^*) + \frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \cdot \mathbf{Q}\mathbf{B}^* \right\} (E_{\theta^*}\mathbf{z} - E_\theta \mathbf{z}).$$

We rewrite the two summands in the statement above as

$$\frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \cdot \mathbf{Q} \right\} \cdot (\mathbf{z}^* - E_{\theta^*}\mathbf{z}^*) = -\frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \right\} \cdot \mathbf{Q} \cdot (\mathbf{z}^* - E_{\theta^*}\mathbf{z}^*) +$$

$$[\phi^T D^{-1}\phi]^{-1} \cdot \frac{\partial}{\partial\theta_j} \left\{ \mathbf{Q} \right\} \cdot (\mathbf{z}^* - E_{\theta^*}\mathbf{z}^*) \frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \cdot \mathbf{Q}\mathbf{B}^* \right\} (E_{\theta^*}\mathbf{z} - E_\theta \mathbf{z}) =$$

$$\frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \right\} \cdot \mathbf{Q}\mathbf{B}^* (E_{\theta^*}\mathbf{z} - E_\theta \mathbf{z}) + [\phi^T D^{-1}\phi]^{-1} \cdot \frac{\partial}{\partial\theta_j} \left\{ \mathbf{Q}\mathbf{B}^* \right\} (E_{\theta^*}\mathbf{z} - E_\theta \mathbf{z}). \tag{48}$$

It suffices to establish that $\forall r > 0\ \exists \mathcal{B}^*_r$ and $N_r \in \mathbf{N}$ s. t.

$$\lim_{n\to\infty} \mathbf{P}^n_{\theta^*} \left[\sup_{\mathcal{B}^*_r} \| \frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \right\} \cdot \mathbf{Q} \cdot (\mathbf{z}^* - E_{\theta^*}\mathbf{z}^*) \| > r \right] = 0, \tag{49}$$

$$\sup_{\mathcal{B}^*_r} \| \frac{\partial}{\partial\theta_j} \left\{ [\phi^T D^{-1}\phi]^{-1} \right\} \cdot \mathbf{Q}\mathbf{B}^* (E_{\theta^*}\mathbf{z} - E_\theta \mathbf{z}) \| < r \quad \forall n \ge N_r, \tag{50}$$

the proof of similar statements for the two remaining expressions that appear in (48) is completely analogous. We establish (50) and (49) with the help of the two lemmas which actually repeat appropriate parts of more general auxiliary theorem 1, section 3. Therefore, their proofs are skipped.

Lemma 4. *Let $\mathcal{G} = \{g_{i,n}(\theta),\ 1 \le i \le n\}_{n=1}^\infty$ be an equicontinuous, uniformly bounded family of functions on a compact set Θ. Then $g(\theta) = \limsup_{n\to\infty} |\frac{1}{n}\Sigma_{i=1}^n g_{i,n}(\theta)|$ is a continuous, bounded function and*

$$\lim_{n\to\infty} \|\ g(\theta) - \sup_{s\ge n} \left|\frac{1}{s}\Sigma_{i=1}^s g_{i,s}(\theta)\right|\ \|_\Theta = 0,$$

i.e. the convergence is uniform in $\theta \in \Theta$.

Lemma 5. *Let $\mathcal{G}$ be as it is defined in Lemma 4. Let $\mathcal{X} = \left\{ \{X_{i,n}\}_{i=1}^n \right\}_{n=1}^\infty$ be a collection of finite sequences of centered random variables with uniformly bounded variances. We further assume that for each $\{X_{i,n}\}_{i=1}^n \in \mathcal{X}$ the elements are mutually uncorrelated: $Cov_n(X_{i,n}, X_{j,n}) = 0$ if $1 \le i < j \le n$. Then*

$$\sup_\Theta \left| \frac{1}{n}\Sigma_{i=1}^n X_{i,n} \cdot g_{i,n}(\theta) \right| \longrightarrow 0 \ \textit{in probability.}$$

Proof of Theorem 8.

We notice that under Assumptions 1.1 and 1.4 the families $\left\{\frac{m}{\alpha m+\beta}\right\}_{m\in\mathbf{Z}_+}$, $\left\{\frac{m^2}{(\alpha m+\beta)^2}\right\}_{m\in\mathbf{Z}_+}$, and $\left\{\frac{m}{(\alpha m+\beta)^2}\right\}_{m\in\mathbf{Z}_+}$ consist of differentiable equicontinuous uniformly bounded functions; the same is true if we replace each function in the listed families with its partial derivative with respect to θ_j, $0 \le j \le 3$. By $\Phi(\theta)$ we denote the 4×4 matrix function obtained from $\phi^T D^{-1}\phi$ by replacing each entry with its absolute value. Since all x_i are uniformly bounded, it follows immediately from (35), (42) and Lemma 4 that

$$\Phi_1(\theta) = \limsup_{n\to\infty} \left\{ \begin{matrix} \frac{1}{n}\cdot \mathbf{Id}_3 & \mathbf{0} \\ \mathbf{0} & \frac{1}{N} \end{matrix} \right\} \cdot \Phi(\theta) \tag{51}$$

is a continuous matrix function on compact Θ and convergence *is uniform in* $\theta \in \Theta$. Thereby, for all n large enough the elements of the 4×4 matrix

$$\Psi(\theta) = \left\{ \begin{matrix} \frac{1}{n}\cdot \mathbf{Id}_3 & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\cdot\left[\phi^T\cdot D^{-1}\cdot\phi\right] \end{matrix} \right\}$$

are uniformly bounded, whence, by Assumption 1.5, the same holds with respect to the elements of the inverse matrix $\Psi^{-1}(\theta)$.

Denote the normalization matrix by $\mathbf{\Xi}$:

$$\mathbf{\Xi} = \left\{ \begin{matrix} \frac{1}{n}\cdot \mathbf{Id}_3 & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\cdot \end{matrix} \right\}. \tag{52}$$

For an arbitrary differentiable nonsingular matrix function $\mathbf{A}$, the derivative of its inverse is given by $\left[\mathbf{A}^{-1}\right]' = -\mathbf{A}^{-1}\cdot\mathbf{A}'\cdot\mathbf{A}^{-1}$, therefore $\dfrac{\partial}{\partial\theta_j}\left\{[\mathbf{\Xi}\cdot\phi^T D^{-1}\phi]^{-1}\right\} = -\Psi^{-1}(\theta)\cdot\left\{\dfrac{\partial}{\partial\theta_j}\Psi(\theta)\right\}\cdot\Psi^{-1}(\theta)$. By replacing the elements of $\frac{\partial}{\partial\theta_j}\Psi(\theta)$ with their absolute values we obtain a result similar to (51) and it follows that, $\forall 0\le j\le 3$, the matrix $\frac{\partial}{\partial\theta_j}\Psi(\theta)$ consists of uniformly bounded elements. We have demonstrated that for all n large enough $[\mathbf{\Xi}\cdot\phi^T D^{-1}\phi]^{-1}$ and $\dfrac{\partial}{\partial\theta_j}\left\{[\mathbf{\Xi}\cdot\phi^T D^{-1}\phi]^{-1}\right\}$ consist of uniformly bounded functions of θ, thereby (49) and (50) both hold, if $\forall r>0$,

$$\lim_{n\to\infty}\mathbf{P}^n_{\theta^*}\left[\sup_{\mathcal{B}^*_r}\|\mathbf{\Xi}\cdot\mathbf{Q}\cdot(\mathbf{z}^*-E_{\theta^*}\mathbf{z}^*)\| > r\right] = 0, \tag{53}$$

and

$$\sup_{\mathcal{B}^*_r}\|\mathbf{\Xi}\cdot\mathbf{QB}^*(E_{\theta^*}\mathbf{z}-E_\theta\mathbf{z})\| < r \quad \forall n\ge N_r. \tag{54}$$

The rest of the proof is a verification of (54) and (53) with the help of Lemma 4 and Lemma 49 and we proceed accordingly.

The ith summand in the first sum in (46) belongs to a family of equicontinuous uniformly bounded functions in $\theta\in\Theta$:

$$\frac{m_i x_i(\theta_0^*-\theta_0+(\theta_1^*-\theta_1)x_i)}{\alpha m_i+\beta} \in \left\{\frac{mx(\theta_0^*-\theta_0+(\theta_1^*-\theta_1)x)}{\alpha m+\beta}\right\}_{\substack{x\in X,\\ m\in\mathbf{Z}_+}};$$

in a similar way each of the remaining three sums can be put into correspondence with a family of equicontinuous uniformly bounded functions that all equal zero when $\theta=\theta^*$. Therefore (54) follows from Lemma 4.

The product $\mathbf{Q}\cdot(\mathbf{z}^*-E_{\theta^*}\mathbf{z}^*)$ in (53) does not change if $\forall 1\le k\le N+N(N+1)/2$ in each row of $\mathbf{Q}$ we multiply the element that corresponds to z_k^* by $Var_{\theta^*}^{\frac{1}{2}}(z_k^*)$ and simultaneously replace $z_k^*-E_{\theta^*}z_k^*$ with $\frac{z_k^*-E_{\theta^*}z_k^*}{\sqrt{Var_{\theta^*}(z_k^*)}}$ in $\mathbf{z}^*-E_{\theta^*}\mathbf{z}^*$. The families of functions that we associate with each of the rows of the new matrix obtained from $\mathbf{Q}$ in this way are all equicontinuous and uniformly bounded thanks to Assumption 1.3 (e.g.

$$\frac{m_i}{2(\alpha m_i+\beta)^2}\in\left\{\frac{m^{\frac{3}{2}}(\theta_0^*-\tilde{\theta}_0+(\theta_1^*-\theta_1)x)\cdot\sqrt{\alpha^*m+\beta^*}}{(\alpha m+\beta)^2}\right\}_{\substack{x\in X,\\ m\in\mathbf{Z}_+}}\cup\left\{\frac{m(\alpha^*m+\beta^*)}{(\alpha m+\beta)^2}\right\}_{m\in\mathbf{Z}_+},1\le i\le n,$$

for the third row), and the vector obtained from $\mathbf{z}^*-E_{\theta^*}\mathbf{z}^*$ contains $P_{\theta^*}^n$-centered mutually uncorrelated random variables with unit variances. Lemma 5 now establishes (53).

Corollary of the Proof. Define $\rho^*=\rho_n^*=\mathcal{U}(\theta^*)-\theta^*$. For any $r>0$,

$$\lim_{n\to\infty}\mathbf{P}_{\theta^*}^n\left[\|\rho^*\|>r\right]=0, \tag{55}$$

by (51) and (53).

Since $\|\mathcal{U}(\theta)-\theta^*\|\le\|\mathcal{U}(\theta)-\mathcal{U}(\theta^*)\|+\|\rho^*\|$, Theorem 8 and (55) imply that there exists an open ball $\mathcal{B}^*$, $\theta^*\in\mathcal{B}^*$, such that the probability of $\mathcal{U}\big|_{\mathcal{B}^*}$ not being a contraction operator tends to zero as $n\to\infty$. The fixed point of $\mathcal{U}\big|_{\mathcal{B}^*}$ satisfies

$$\phi^TD^{-1}(\mathbf{z}-E\mathbf{z})=0. \tag{56}$$

In view of (35), (56) is equivalent to

$$\phi_{\tilde{\mathbf{B}}\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})D_{\tilde{\mathbf{B}}\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})(\tilde{\mathbf{z}}-E\tilde{\mathbf{z}})=0, \tag{57}$$

where $\tilde{\mathbf{z}}$ is given by (36) and $\tilde{\mathbf{B}}$ by (38). The matrix $\phi_{\tilde{\mathbf{B}}\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})D_{\tilde{\mathbf{B}}\mathbf{z}+\tilde{\mathbf{t}}}(\tilde{\theta})$ is given explicitly by (41) and it is a straightforward computation to verify that (57) is equivalent to

$$\nabla\mathbf{F}\cdot V^{-1}\cdot\nabla^T\mathbf{F}\cdot\begin{Bmatrix}\theta_0\\ \theta_1\end{Bmatrix}=\nabla\mathbf{F}\cdot V^{-1}\mathbf{y},$$
$$\left\{tr\left\{V^{-1}\mathbf{U}\mathbf{U}^T\right\}(\mathbf{y}-\mathbf{F})^TV^{-1}\mathbf{U}\mathbf{U}^TV^{-1}(\mathbf{y}-\mathbf{F})\right\},$$
$$\left\{tr\left\{V^{-1}\right\}=(\mathbf{y}-\mathbf{F})^TV^{-2}(\mathbf{y}-\mathbf{F})\right\},$$

$\mathbf{F}$ is a function of two variables: θ_0 and θ_1 (see (29)). The equations above are identical to the likelihood equations presented on p.749 in [36].

12.5. LAN and LAM

Next we establish the Local Asymptotic Normality (LAN) and, in addition, prove the *uniform convergence to zero in $P_{\theta^*}^n$ - probability* of the residual term in the asymptotic expansion. Throughout this section we suppose that Assumptions 1.1–1.7 hold.

Theorem 9. *Fix an arbitrary compact set $\mathbf{T}\subset\mathbf{R}^4$ containing zero. Set*

$$L_n(u)=\ln[dP_{\theta^*+\Xi^{1/2}\cdot u}^n/\ d\ P_{\theta^*}^n],\ u\in\mathbf{T},$$

where Ξ is given by (52).

Then the following decomposition (uniform LAN) holds:

$$L_n(u) = \lambda^T u - (1/2)u^T J u + \psi_n(u), \tag{58}$$

where

$$\lambda \sim \mathcal{N}(0, J), \quad J = \left\{ \begin{matrix} 1/\alpha^* & \bar{x}/\alpha^* & 0 & 0 \\ \bar{x}/\alpha^* & \overline{x^2}/\alpha^* & 0 & 0 \\ 0 & 0 & 1/2\alpha^* & 0 \\ 0 & 0 & 0 & 1/2\beta^* \end{matrix} , \right\}$$

and

$$\lim_{n\to\infty} P^n_{\theta^*}\left[\sup_{u\in\mathbf{T}} |\psi_n(u)| > \rho\right] = 0 \quad \forall\rho > 0. \tag{59}$$

Remark 6. The LAN itself only requires that the residual term in (58) converges to zero pointwise.

Proof. We use the multivariate version (valid for twice continuously differentiable functions) of the expansion that was first published in 1797 by J.L.Lagrange:

$$F(x) = F(a) + \nabla F(a)(x-a) + \frac{1}{2}(x-a)^T \nabla^2 F(a)(x-a) +$$
$$(x-a)^T \cdot \Big\{ \int_0^1 (1-t)\left[\nabla^2 F(a + t(x-a)) - \nabla^2 F(a)\right] dt \Big\} \cdot (x-a), \tag{60}$$

where $\nabla F(x)$ is the gradient of $F(x)$ and $\nabla^2 F(x) = \| \frac{\partial^2}{\partial x_i \partial x_j} F(x) \|_{i,j}$.

The proof of Theorem 9 is based on applying the above decomposition to $L_n(u)$ when $a = 0$ (note that $L_n(0) = 0$).

We begin by demonstrating that the first two terms in (58) are the limits of $\nabla L_N(0)u$ and $(1/2)u^T \nabla^2 L_N(0)u$ respectively; we then represent the residual term in an integral form (cf. (60)) and observe that (59) follows from a stronger statement which we prove with the help of Lemmas 4 and 5: $\forall\rho > 0$, $\forall i, j$ such that $0 \le i, j \le 3$,

$$\lim_{r\to 0}\limsup_{n\to\infty} P^n_{\theta^*}\left[\sup_{\theta\in\mathcal{B}_r(\theta^*)} \left| \Big(\nabla^2 L_n(\Xi^{-\frac{1}{2}}[\theta-\theta^*]) - \nabla^2 L_n(0)\Big)_{i,j} \right| > \rho \right] = 0, \tag{61}$$

here $\mathcal{B}_r(\theta^*)$ is a ball of radius r centered at θ^*, $\left(\nabla^2 L_n(u)\right)_{i,j} = \partial^2/\partial\theta_i\partial\theta_j L_n(u)$. Prior to proceeding in accordance with this plan we pause to prove the following useful result, which can also be found elsewhere:

Lemma 6. *Let* $\xi = (\xi_1, \ \ldots \ , \xi_n)^T \sim \mathcal{N}(0, V)$. *Then*

$$Var[tr\left\{\xi \cdot \xi^T\right\}] = 2 \cdot tr\left\{V^2\right\}. \tag{62}$$

Proof of the Lemma. When V is diagonal (i.e. when the components of ξ are mutually independent), (62) is established by direct computation. The general case follows since for any orthogonal O and an arbitrary $n \times n$ matrix A we have $tr\left\{O^T \cdot A \cdot O\right\} = trA$.

Let V, just like in (31), denote $Cov_{n,\theta}(\mathbf{y})$ and set $G = G_n = \mathbf{U}\mathbf{U}^T = diag(\mathbf{1}_{m_1} \cdot \mathbf{1}^T_{m_1}, \ldots, \mathbf{1}_{m_n} \cdot \mathbf{1}^T_{m_n})$, then $VG = GV$ and $H_i^2 = m_i H_i$ $(H_i = \mathbf{1}_{m_i} \cdot \mathbf{1}^T_{m_i})$. In this notation, $\forall\theta \in \Theta$

$$dP^n_\theta = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(\mathbf{y}-\mathbf{F})^T V^{-1}(\mathbf{y}-\mathbf{F})} dy_1 \ldots dy_N. \tag{63}$$

Throughout the remainder of the proof we apply repeatedly the following well-known identities that hold for any two non-singular differentiable matrix functions A and B such that AB and BA are both defined:

$$d\left\{A^{-1}\right\} = -A^{-1} \cdot \{dA\} \cdot A^{-1},\, d\ln|A| = tr\left\{A^{-1} \cdot dA\right\},\, tr\left\{AB\right\} = tr\left\{BA\right\}. \tag{64}$$

The above identities combined with (63) allow us to list the components of $\nabla L_n(0)$:

$$\frac{\partial}{\partial\theta_0}L_n(0) = \frac{1}{\sqrt{n}}\mathbf{1}_N^T \cdot V^{*-1} \cdot (\mathbf{y} - \mathbf{F}^*),\ \frac{\partial}{\partial\theta_1}L_n(0) = \frac{1}{\sqrt{n}}\mathbf{x}^T \cdot V^{*-1} \cdot (\mathbf{y} - \mathbf{F}^*),$$
$$\frac{\partial}{\partial\theta_2}L_n(0) = -\frac{1}{2\sqrt{n}}\left\{(\mathbf{y} - \mathbf{F}^*)^T V^{*-1} G V^{*-1}(\mathbf{y} - \mathbf{F}^*) - tr(GV^{*-1})\right\},$$
$$\frac{\partial}{\partial\theta_3}L_n(0) = \frac{1}{2\sqrt{N}}\left\{(\mathbf{y} - \mathbf{F}^*)^T V^{*-2}(\mathbf{y} - \mathbf{F}^*) - trV^{*-1}\right\}, \tag{65}$$

where $\mathbf{x} = \frac{\partial}{\partial\theta_1}\mathbf{F}\big|_{\theta^*}$. All expressions in (65) are centered random variables. Our analysis of the asymptotics of $\nabla L_n(0) \cdot u$ starts with a verification of $\lim\limits_{n\to\infty} Cov_{\theta^*}(\nabla^T L_n(0)) = J$. Here we resort essentially to the same techniques employed in section 12.4.

Let $\mathcal{O}$ be the orthogonal map that diagonalizes V (see (31) and (32)), then

$$Cov_{\theta^*}(\frac{\partial}{\partial\theta_0}L_n(0), \frac{\partial}{\partial\theta_1}L_n(0)) = \frac{1}{n}E_{\theta^*}tr\left\{(\mathcal{O}^T\mathbf{1}_N)^T\mathcal{O}^T V^{*-1}\mathcal{O}\mathbf{y}^*\mathbf{y}^{*T}\mathcal{O}^T V^{*-1}\mathcal{O}(\mathcal{O}^T\mathbf{x})\right\} =$$

$$\frac{1}{n}tr\left\{(\mathcal{O}^T\mathbf{1}_N)^T\mathcal{O}^T V^{*-1}\mathcal{O}(\mathcal{O}^T\mathbf{x})\right\} = \frac{1}{n}\Sigma_{i=1}^n\frac{x_i}{\alpha^* + \beta^*/m_i} \to \frac{\bar{x}}{\alpha^*}$$

as $n \to \infty$.

In a similar fashion one can show that $Cov_{\theta^*}(\frac{\partial}{\partial\theta_2}L_n(0), \frac{\partial}{\partial\theta_3}L_n(0)) \to 0$ as $n \to \infty$.

To compute variances for $\frac{\partial}{\partial\theta_2}L_n(0)$ and $\frac{\partial}{\partial\theta_3}L_n(0)$ we apply Lemma 6 to Gaussian vectors $G^{\frac{1}{2}}V^{*-1}(\mathbf{y} - \mathbf{F}^*)$ and $V^{*-1}(\mathbf{y} - \mathbf{F}^*)$ respectively:

$$Var_{\theta^*}[\frac{\partial}{\partial\theta_2}L_n(0)] = \frac{1}{2n}tr\left\{GV^{*-1}GV^{*-1}\right\} \text{ and } Var_{\theta^*}[\frac{\partial}{\partial\theta_3}L_n(0)] = \frac{1}{2N}tr\left\{V^{*-2}\right\},$$

which go to $\dfrac{1}{2\alpha^*}$ and $\dfrac{1}{2\beta^*}$ as $n \to \infty$. After careful inspection of the remaining terms one concludes that J is indeed the limit of the covariance matrix of the vector $\nabla^T L_n(0)$. Set $\mathbf{y}^* = \mathcal{O}^T \cdot (\mathbf{y} - \mathbf{F}^*)$. We construct a collection of sequences of mutually independent random vectors $\left\{\{\xi_{n,i}\}_{i=1}^n\right\}_{n=1}^\infty$,

$$\xi_{n,i} = Cov_{\theta^*}^{-\frac{1}{2}}\left(\nabla^T L_n(0)\right) \times$$

$$\left\{\begin{matrix} \frac{\sqrt{m_i}y^*_{m_1+\ldots+m_i}}{\sqrt{n}(\alpha^* m_i+\beta^*)} \\ \frac{\sqrt{m_i}x_i y^*_{m_1+\ldots+m_i}}{\sqrt{n}(\alpha^* m_i+\beta^*)} \\ \frac{1}{2\sqrt{n}}\frac{m_i}{\alpha^* m_i+\beta^*}\left(\frac{y^{*2}_{m_1+\ldots+m_i}}{\alpha^* m_i+\beta^*} - 1\right)\frac{1}{2\sqrt{N}}\left(\frac{1}{\alpha^* m_i+\beta^*}(\frac{y^{*2}_{m_1+\ldots+m_i}}{\alpha^* m_i+\beta^*} - 1) + \frac{1}{\beta^*}\Sigma_{l_{i-1}<k<l_i}(\frac{y^{*2}_k}{\beta^*} - 1)\right) \end{matrix}\right\},$$

$1 \le i \le n$, $l_i = m_0 + \ldots + m_i$ and $m_0 = 0$ by definition. Since $\forall n\ \Sigma_{i=1}^n\xi_{n,i} = Cov_{\theta^*}^{-\frac{1}{2}}\left(\nabla^T L_n(0)\right) \cdot \nabla^T L_n(0)$, a modification of Lindeberg-Feller's Theorem implies that $\nabla^T L_n(0) \to \mathcal{L}\left(\mathcal{N}(0, J)\right)$. Then by the Kolmogorov's famous lifting theorem (on existence of a stochastic processes), $\nabla^T L_n(0)\cdot u \to \lambda^T u$, $\lambda \sim \mathcal{N}(0, J)$, and we are done with the first term.

The components of $\nabla^2 L_n(0)$ are treated similarly. By Lemma 6 the variances of

$$\frac{\partial^2}{\partial\theta_2^2}L_n(0) = \frac{1}{2n}\left[tr\left\{GV^{*-1}GV^{*-1}\right\} - 2(\mathbf{y}-\mathbf{F}^*)V^{*-1}GV^{*-1}GV^{*-1}(\mathbf{y}-\mathbf{F}^*)\right]$$

tend to zero as $n \to \infty$, thereby they converge in $P^n_{\theta^*}$-probability to $-1/2\alpha^*$ and $-1/2\beta^*$, which are the limits of their expectations. Likewise we notice that the variances of the centered expressions

$$\frac{\partial^2}{\partial\theta_0\partial\theta_2}L_n(0) = -\frac{1}{n}\mathbf{1}_N^T V^{*-1}GV^{*-1}(\mathbf{y}-\mathbf{F}^*),$$

$$\frac{\partial^2}{\partial\theta_0\partial\theta_3}L_n(0) = -\frac{1}{\sqrt{nN}}\mathbf{1}_N^T V^{*-2}(\mathbf{y}-\mathbf{F}^*),$$

$\dfrac{\partial^2}{\partial\theta_1\partial\theta_2}L_n(0)$ converge to zero.

The rest of the matrix elements of $\nabla^2 L_n(0)$ can be computed directly by applying orthogonal transformation $\mathcal{O}$, and it follows that $(1/2)u^T\nabla^2 L_n(0)u$ converges to $-(1/2)u^T J u$, hence we are done with the second term.

It is only left to verify (61), which is done by repeated applications of Lemmas 4, 5, and 6. The proof of (61) is carried out in complete analogy with the proof of (53)–(54) and the analysis done for $\nabla^2 L_n(0)$ above and is therefore omitted.

To prove (58) we need the following strengthening of (55) (where $J^{1/2}_{\theta^*}$ is defined in the statement of Theorem 9

$$\lim_{n\to\infty}\mathbf{E}^n_{\theta^*}\left[f(J^{1/2}_{\theta^*}\boldsymbol{\Xi}^{-1/2}\rho^*)\right] = \int f(u)(2\pi)^{-2}e^{-|\mathbf{u}|^2}d\mathbf{u} \tag{66}$$

for any continuous bounded function $f(\cdot)$ on $\mathbf{R}^4$ (weak convergence to the standard normal distribution). This statement follows from multidimensional Lindeberg-Feller theorem [42] and the computations of section 12.4.

The LAM property of θ^1 for bounded loss functions uniformly over qualified initial guesses θ^0 follows from theorem 3.4.2 of [28], we outline here the main steps of its derivation skipping the more involved proof of the convergence of moments of any order to the moments of the limiting distribution (which is also true for our gaussian model). We have

$$\theta^1 - \theta^* = \mathcal{U}(\theta^0) - \theta^* = \rho^* + R(\theta^0).$$

Set $\mathbf{A}_n = [\phi^T D^{-1}\phi]^{-1}\phi^T D^{-1}$. Then by the integral Taylor expansion of the first order (cf. (60))

$$\boldsymbol{\Xi}^{-1/2}R = R_1 + R_2,$$

$$R_1 = \boldsymbol{\Xi}^{-1/2}\int_0^1 \frac{\partial}{\partial\theta}\left\{\mathbf{A}_n\cdot(\mathbf{z} - E^*\mathbf{z})\right\}|_{\theta^*+t(\theta^0-\theta^*)}\cdot(\theta^0-\theta^*)dt,$$

$$R_2 = \boldsymbol{\Xi}^{-1/2}\mathbf{A}_n(\theta^0)\cdot\int_0^1\left[\phi(\theta^0+t(\theta^*-\theta^0)) - \phi(\theta^0)\right]\cdot(\theta^*-\theta^0)dt.$$

Taking into account the boundedness in probability of $\boldsymbol{\Xi}^{-1/2}(\theta^0-\theta^*)$, we obtain:

$$\lim_{n\to\infty} \mathbf{P}^n_{\theta^*}\left[\sup_{\mathcal{B}^*_r} \|R(\theta^0)\| > r\right] = 0.$$

This estimate and (66) imply (27) since the maximized expression in the left-hand side of the latter does not depend on $\theta^* \in \Theta$, whereas the residual terms are uniformly negligible.

12.6. Asymptotically optimal Design

The inverse covariance matrix J of the limiting distribution for the vector

$$([\sqrt{n}(\hat{\beta}_0 - \beta), (\hat{\beta}_1 - \beta_1), (\hat{a} - a)], \sqrt{N}(\hat{b} - b))$$

is given at the line just below (58). We see that it does not depend on the mutual relationship between $m_1, \ldots, m_I$ as long as $\min\{m_1, \ldots, m_I\} \to \infty$ and Assumption 1.7 holds. Particularly, the part of J corresponding to the parameters θ_0, θ_1 is the same as for n independent homoskedastic observations of the straight line regression model at the points $x_i, i = 1, \ldots, I$.

Remark 7. Our asymptotic design results differ drastically from those for ANOVA models with deterministic parameters ([18]), where balancedness plays the key role.

Remark 8. Our asymptotic theory can be easily generalized to the case of general linear model for the mean $X\gamma$ of our mixed model, where

$$X = \left\{\begin{matrix} f_1(x_1) \ldots f_p(x_1) \\ \ldots \quad \ldots \quad \ldots \\ f_1(x_n) \ldots f_p(x_n) \end{matrix}\right\}, \quad f_i : X \to R^1,$$

are continuous functions on a compact subset X of $\mathbf{R}^q$.

Remark 9. Designs optimizing a convex differentiable function $\Phi(J^{-1})$ by choosing $x_1, \ldots, x_I$ can be easily found (at least numerically) by the standard methods known for homoskedastic independent measurements (see, e.g. [8].

13. MULTISAMPLE MIXTURE PARAMETER ESTIMATION

13.1. Introduction

This section was applied to a study of size and age structure of a fish population on the basis of length-frequency analysis. Because of a short reproductive season in cold and temperate waters, the fish population may be considered as a mixture of discrete generations. Each generation has usually a unimodal distribution of body length. Therefore, a length-frequency histogram for a sample from the population may be treated as the one obtained from a mixture of the component distributions. Field data usually consist of about 100 samples. Samples from distinct locations within the area of population are not homogeneous spatially because the generations are poorly mixed within the area. For example, young generations usually predominate in coastal waters while mid-aged generations predominate in deeper waters.

A closely related application of the model occurs in latent structure analysis, particularly in the social science (Goodman (1974)). A similar problem was considered by Skene (1978), who studied the sets of data, each from a different group of patients, assuming that the component densities are the same for all groups but that the sets of mixing weights may differ from group to group. Maximum likelihood estimation using the EM algorithm was suggested. However, asymptotic properties of the estimates were not discussed. It is unclear whether the approach based on estimation of weights of classes in each sample provides consistent parameter estimates because the number of classes weights is increasing proportionally with the number of samples.

13.2. Method

Let us consider N samples. Each sample is represented by a histogram. Let J-dimensional vector Y_i describe the i-th histogram of J group intervals as a mixture of K components, its j-th element is

$$y_{j,i} = \sum_{s=1}^{K} p_{j,s}(\theta)c_{si} + u_{j,i}, \ j = 1, \dots, J, \ i = 1, \dots, N.$$

Here c_{si} is the weight of the s-th component density in i-th histogram, $p_{j,s}, s = 1, \dots, K$, is the multinomial probability depending on unknown $\theta \in \mathbf{R}^M$; $u_{j,i}$ is a random error. The vector form of the above formula is the following

$$Y_i = P(\theta)C_i + u_i, \ i = 1, \dots, N,$$

where $C_i = (c_{si}, s = 1, \dots, K)$, and the matrix $P(\theta)$ with entries $p_{j,s}(\theta), j = 1, \dots, J, s = 1, \dots, K$, is assumed to have rank $J - 1$.

Under fixed C_i, the mean $E(u_i|C_i) = 0$, and its conditional covariance is

$$Cov(u_i|C_i) = 1/n_i[diag(P(\theta)C_i) - (P(\theta)C_i)(P(\theta)C_i)^T],$$

where numbers n_i of observations the i-th histogram is based on are assumed to be independent identically distributed (i.i.d.) random with mean n (independent of $u_i, i = 1, \dots, N$). It is straightforward that $Cov(u_i|C_i)\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is the vector with all components 1.

From now on consider $C_i, i = 1, \dots, N$, as K-dimensional *i.i.d. multinomial random variables with the common mean* $\pi(\alpha) = E(C_i), \alpha \in \mathbf{R}^q, q < K$, *describing the relative proportions of generations* (in the framework of the example from the fishery research studied by us). The covariances of C_i are

$$D_i = \left[diag(\pi) - \pi\pi^T \right].$$

Thus, we suppose that all the weight vectors $C_i, i = 1, \dots, N$, are drawn from the same multinomial distribution with K classes. This assumption provides an opportunity to study asymptotic properties of the model depending on a finite number of real-valued parameters $\beta^T = (\theta^T, \alpha^T)$.

As compared to the related approach of [45], the ages of populations correspond to the latent classes, whereas his assumption on independence of symptoms from the latent classes is weakened by our multinomial distribution assumption. Instead of Bayesian scheme we study estimation in the parametric model.

Denoting $r_i = C_i - \pi$, we obtain the multivariate regression model depending on unknown parameters $\beta^T = (\theta^T, \alpha^T)$:

$$Y_i = P(\theta)\pi(\alpha) + e_i, \tag{67}$$

where $e_i = P(\theta)r_i + u_i$ are independent random vector-errors with

$$E(e_i) = 0, \tag{68}$$

and $W_i := Cov(Y_i)$ is given by

$$(1 - 1/n_i)P(\theta)[diag(\pi) - \pi\pi^T]P^T(\theta) + n_i^{-1}diag[P(\theta)\pi] - P(\theta)\pi\pi^T P^T(\theta). \tag{69}$$

The last expression and the equality $W_i\mathbf{1} = \mathbf{0}$ follow from the identity

$$Cov(Y_i) = ECov(Y_i|C_i) + CovE(Y_i|C_i),$$

and the mutual independence of vectors u_i, r_i and n_i.

The equations (67) to (69) constitute a multivariate regression model. *Parameters β are assumed to be identifiable given the common mean.* Singular covariances of vector-observations depend on unknown parameters. This is an example of Q-model studied in sections 3–9. The REAGAN algorithm with the weights taken as a smooth g-inverse of the covariance matrix (69) in Q-model possesses the same optimal asymptotic properties as the optimally weighted LS estimator with the known covariance structure (section 3). The t-th step of the REAGAN consists of finding the best linear unbiased estimate $\beta^t - \beta^{t-1}$ for the weighted LS fitting residuals of the previous approximation $Y_i - P(\theta^{t-1})\pi(\alpha^{t-1})$ by $F(\beta^{t-1})(\beta^t - \beta^{t-1}), i = 1, \dots, N$, where $F(\beta) = \partial P(\theta)\pi/\partial\beta$ and the weight matrices are the generalized inverses (g-inverses) W_i^- for W_i satisfying the equation

$$W_i W_i^- W_i = W_i,$$

see for example [40], section A.12. According to their theorem A.77, (where their *nonsingular* should be replaced with *singular* to correct the obvious typo) the matrix $\Omega_i = (W_i + \mathbf{1} \times \mathbf{1}^T)^{-1}$ is a smooth g-inverse to W_i because $\mathbf{1}$ is the only vector in the null space of W_i. Putting Ω_i as weight matrices of the multivariate REAGAN, we get a desired estimate $\hat{\beta} = \lim \beta^t$ (as $t \to \infty$) of the true value β^* for the parameter β. It is proved in section 3, that this limit exists in Probability as $N \to \infty$, it is consistent for the initial guess being sufficiently close to β^*, and $\sqrt{N}(\hat{\beta} - \beta^*)$ has asymptotically normal distribution with zero mean and covariance matrix

$$A = [\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} F^T(\beta^*)\Omega_i F(\beta^*)]^{-1}.$$

Remark 10. If the distribution of n_i depends on N in such a way that $n_i \to \infty, i = 1, \dots, N$, as $N \to \infty$, then the normalization rates are different for θ and β, cf. section 12.

Remark 11. If $rankP(\theta) < J - 1$, we can still find a smooth g-inverse to W_i for running the REAGAN, using the method developed in [39]. Namely, it is enough to compose a matrix H with columns constituting the base in the null space of $P(\theta)$. Then $(W_i + HH^T)^{-1}$ is a g-inverse to W_i according to [39], statement 2 of 4a.3.

If the initial guess β^0 is $\sqrt{N}$-consistent, then the first iteration β^1 is already asymptotically normal with the same parameters. It should be emphasized that our method of estimation based on fitting two first moments of distribution often uses asymptotically all information on unknown parameters contained in the data, particularly for a regular curved multinomial (and in general for any regular curved exponential family of distributions) (see our section 10). Namely, the estimate $\hat{\beta}$ asymptotically approaches the maximum likelihood estimate in the case, when the family of distributions depending on β is a regular curved exponential one.

14. DENSITY PARAMETER ESTIMATION FROM STRATIFIED SAMPLE

Our schematic sketch of parameter estimation from stratified sample follows [27]. Suppose the range of continuous univariate distribution density $f((x), \theta)$ w.r.t the Lebesgue measure dx is split into many strata such that the limiting smooth normalized distribution of the repeated sample ν_i^N from stratum (bin) S_i^N also converges to $f(\cdot), \theta$ together with its smooth derivative w.r.t. θ. Occupancy numbers ν_i^N follow Multinomial distribution with means $\mu_i^N = \int_{S_i^N} f(x)dx$ and covariance function

$$V(\theta) = N^{-1}[diag\{\mu^N\} - \mu \times (\mu^N)^T].$$

We use the same g-inverse to $V(\theta)$ as in section 13 and define the REAGAN procedure accordingly:

$$\theta^{s+1} = \Theta_{\nu^N}(\theta^s) = (\Phi^T V^- \Phi)\Phi^T V^-(\nu^N - \mu^N),$$

matrix $\Phi = \partial\mu^N(\theta)/\partial\theta$.

Proof of the REAGAN convergence and asymptotic normality in [27] uses the $\sqrt{\nu^N}$-transformation of occupancy numbers which converts their covariance matrix asymptotically into

$$Q_N = (\mathbf{I} - \sqrt{\mu^N}\sqrt{\mu^N}^T).$$

Operator Q_N projects ν_i^N orthogonally to vector $\sqrt{\mu^N}$, $Q_N^2 = Q_N$, $||Q_N|| = 1$. This means informally that we work with the uniform distribution on the intersection of the circumference with the positive octant after this transformation. The same transformation converted the two-allelic R. Fisher's genetic drift asymptotically into an isotropic Brownian motion on the same part of the circumference. This enabled the R. Fisher's simplified asymptotic analysis, see e.g. [24].

In the limit to an infinitesimal stratification, the covariance matrix approaches the Fisher's lower bound

$[\int \partial\sqrt{f(\theta)}\partial\sqrt{f^T(\theta)}]^{-1}/N$ confirming the asymptotic efficiency of our estimator.

15. M-ESTIMATES AND SMALL ERROR CASES

15.1. Δ-method

The class of gaussian Q-models with *small noise* $\gamma\delta, \gamma \to 0, Cov\delta = D$, is obviously asymptotically **closed w.r.t. smooth transformations** $g(\cdot)$ of responses y due to the well-known **Δ-method** [39].

Namely, the principal term of the mean of transformed gaussian approximation is $g(\mu)$, while the principal in γ term of the Covariance is $\gamma^2\partial g D\partial^T g$ which is also a Q-model.

15.2. M-estimates

Let us point out that **REAGAN** for a Q-model can be easily made more **robust**, i.e. ignoring or almost ignoring a certain percentage of outliers and having a higher breakdown point. For this purpose, it is sufficient to multiply the weight matrix of our Q-model by a factor which is decreasing as $\|y_i - \mu_i(\theta)\| \to \infty$. Despite the fact that now we can not consider the weight as $Cov^{-1}[y_i]$, the method of proving asymptotic properties of estimators stays the same. The efficiency of such procedures has been studied in detail in the theory of M-estimators[14].

15.3. Small Errors in Controllable Explanatory Variables

The classical Linear Models (LM) deal with precisely known x-predictors. Under more realistic admission of small unbiased random errors in controllable predictors, the LS method is no more consistent due to additional bias and variance of y-observations depending on the slopes of regressors. An appropriate Q- model and REAGAN iterative estimation saves the situation.

Early references are [4, 7], where consistent suboptimal algorithms were constructed without rigorous convergence of iterations proof.

A more rigorous [49] considers observations series such that the corrections' order of magnitude is possible to estimate for finite samples.

Given m-vector θ_* of non-random parameters and N-vectors $\mathbf{y}, h, \delta$ of observations as well as independent random errors in predictors and responses $\eta(u_i, \theta_*)$ with twice continuously differentiable $\eta(\cdot)$, consider model

$$y_i = \eta(u_i + h_i, \theta) + \delta_i, i = 1, \ldots, N,$$

where u_i, h_i are respectively fixed and random p-vectors, $Eh \equiv 0, E\delta \equiv 0$, the normalized design ε_N of predictors weakly converges to measure ε such that the limiting information matrix $M(\theta_*)$ *introduced further by (70) is non-singular.*

Also, assume that $\sqrt{\Gamma} N^{1/6} h = \nu$ converges to a limiting non-degenerate distribution which does not depend on N and has four finite moments.

The second order Taylor decomposition gives (denoting second derivative over xx as η_{xx}, etc.):

$$Ey_i = \eta(u_i, (\sigma^2 + \partial\eta(u_i, \theta_*)^T \partial\eta(u_i, \theta_*)\Gamma N_*^{-1/3}) + \eta_{xx}(u_i, \theta_*)/2 + O(N^{-1}),$$

$$Var(y_i) = \sigma^2 + \partial\eta(u_i, \theta_*)^T \partial\eta(u_i, \theta_*)\Gamma N^{-1/3} + O(N^{-2/3}.)$$

Introduce

$$\psi(u_i, \theta) = \eta(u_i, \theta) + \eta_{xx}(u_i, \theta)/2 + O(N^{-1}),$$

$$\lambda^{-1} = \sigma^2 + \partial\eta(u_i, \theta)^T \partial\eta(u_i, \theta)\Gamma N^{-1/3},$$

$$F^T(u, \theta, \gamma) = [\partial\psi_\theta, \partial\psi_\gamma],$$

$$M(\theta, \gamma) = \sum F(u_i, \theta, \gamma) F^T(u_i, \theta, \gamma)\lambda(u_i, \theta, \gamma). \tag{70}$$

Consider a REAGAN iterative algorithm $\mathcal{A}_y(\cdot)$

$$(\theta^{s+1}, \gamma^{s+1}) = \mathcal{A}_y(\theta^s, \gamma^s),$$

where operator $\mathcal{A}_y$ means Argmin $\sum_1^N (y_i - \psi(u_i, \theta, \gamma)^2 \lambda(u_i, \theta^s, \gamma^s)$.

The methods described in sections 3–9 enable proof in [49] of the convergence in Probability, asymptotic normality and local Minimaxity of estimates in the class of bi-linear updates via $\mathcal{A}_y(\cdot)$.

16. DISCUSSION

A sample of applications displayed above shows that Quasilinear models constitute a flexible broad extension of Linear models admitting an effective analysis.

Much more restrictive Generalized Linear models [37], when applicable, construct a bridge between the REAGAN estimates for Q-models and iterative methods of Maximum Likelihood.

REFERENCES

1. Anderson R. L. (1975). Designs and estimators for variance components. *J.N. Srivastava ed., A Survey of Statistical Design and Linear Models*, North Holland Publ. Co, (1975), 1–29.

2. Barndorf-Nielsen O. *Information and Exponential Families*, Wiley, N.Y., 1978.

3. Ben-Israel A., Greville Th. *Generalized inverses. Theory and applications (2nd ed.).* New York, NY: Springer, 2003.

4. Bergson J. Are there two regressions? *J. Amer. Statist. Assoc.*, **50**, 1950, 166–180.

5. Demidenko E. Z. *Linear and nonlinear regression*, Finance and Statistics, M. 1981 (in Russian).

6. Fedorov V.V. *Theory of optimal experiment*, Nauka, M., 1971 (in Russian).

7. Fedorov, V.V. Regression problems with controllable variables subject to error, *Biometrika*, **61**, 49-56.

8. Fedorov V. V. and Leonov S. L. *Optimal design for nonlinear response models*, CRC Press, 2013.

9. R. A. The correlation between relatives ... *Trans. R. Soc. Edinburgh*, **52**, 1918, 399–433.

10. Fisher R. A. *Statistical methods for research workers*, Oliver and Boyd, London, 1925.

11. Goodman, L.A. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231, 1974.

12. Goldstein H. (1986). Multilevel mixed linear models analysis using iterative generalized least squares. *Biometrika*, **73: 1**, 1986, 43–56.

13. Hartley H. O. and Rao J. N. K. MLE for the mixed ANOVA model. *Biometrika*, **54**, 1967, 93–108.

14. Huber P. (1980) *Robust Statistics*, Wiley, N.Y., 1980.

15. Ibragimov I. A. and Khasminski R. Z. *Statistical estimation. Asymptotic theory*, Springer, N.Y., 1981.

16. Ivanov A. V. An asymptotic expansion for the distribution of the least squares estimator of the nonlinear regression parameter, *Theory Probab. Appl.*, **21**, No.3, 1976, 557-570.

17. Jennrich R. Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.*, **40** 1969, 633-643

18. Kiefer J. Optimum Experimental Designs, *Journal of the Royal Statistical Society, Series B.* **21**, 1959, 272–319.

19. Kushnir A. F. *Statistical and numerical methods of seismic monitoring*, URSS, 2012

20. Le Cam L. *Notes on asymptotic methods in statistical decision theory*, Centre de Recherches Math., U. of Montreal, 1974.

21. Luanchi M. *Asymptotic investigation of iterative estimates* (thesis). Moscow Lomonosov University, Depart. of Math. and Mech, 1983.

22. Magnus J.R. *Linear Structures.* Griffin, London, 1988.

23. Malinvaud E. The Consistency of Nonlinear Regressions, *Ann. Math. Statist.* **41**, 1970, 956-969.

24. Malyutov M., Passekov V., and Rychkov Yu. On the reconstruction of evolutionary trees of human populations resulting from random genetic drift. In: *Weiner, J.S., and J. Huizinga (eds.), The Assesment of Population Affinities in Man*, Clarendon Press, Oxford, 1972, 48–71.

25. Malyutov M. B. On joint square-iterative estimation of the mean and variance components, *Theory Probab. and Appl.*, **26**, No.1, 1981.

26. Malyutov M. B. Asymptotics and applications of iterated reweighted Gauss-Newton algorithm, *Stochastic processes and Applications, ed. by G.I.Ivchenko*, Moscow College of Electronics Press, 1982 (In Russian).

27. Malyutov M. B. and Luanchi M. Estimation of a parametric density function using quantified samples, *Statistical Methods*, Perm University Press, 1983, English translation is *Journal Soviet Mathematics* V. 41(1988), No 1.

28. Malyutov M. B. Lower Bounds for the Mean Duration of Sequentially Programmed Experiments, *Sov.Math.(Izv.Vuzov)*, 1983,**27**, no.ll (translated into English by Allerton Press), 21-47.

29. Malyutov M. B, Horna Uaraka L. A. and Spokoiny V. G. *Proceedings of the Steklov Institute of Mathematics*, 1994, **202**, 155–168. (Russian version) *Trudy MIAN*, **202**, 1993, 190–208.

30. Malyutov M. B. and Protassov R. Functional approach to the asymptotic normality of the non-linear least squares estimator, *Statistics and Probability Letters*, **44**, No. 4, 1999, 409-416.

31. Malyutov, M.B. and Protassov, R.S. LAN and LAM, Convergence of Iterative Estimates and Optimal Design in Gaussian One-Way Model, *Journal of Statistical Planning and Inference*, **100(2)**, 249-279, 2002.

32. Malyutov M. B. Local Asymptotic Normality in mixed Gaussian ANOVA model, Proceedings of the first Bernoulli congress. Tashkent, **2**, VNU Publishers, 1986.

33. Malyutov M. B. and Stolyarenko D. A. Fitting Multisample Multinomial Mixture Model, *American Journal of Mathematical and Management Sciences*, **21**, 101–107, 2001.

34. Malyutov M. B., Nikiforov A. N.. and Yukhananov R. Robust rewighted orthogonal regression for microarray normalization, *Proceedings of 4th International Conference on Health Informatics and Bioinformatics (HIBIT 2009), Ankara, April 16-17, 2009*, 2009.

35. Matos M. R. A. *Asymptotic estimation theory of nonlinear regression parameters, PhD*, Moscow State University, 1987.

36. Miller J. J. Asymptotics for MLE's in the mixed ANOVA model, *Ann. Statist.*, **5**, 1977, 746-762.

37. McCullagh P. and Nelder J. A. *Generalized Linear Models*, 2nd ed. Chapman and Hall/CRC Press, 1989.

38. Romanovsky V. I. Mathematical Statistics, v.2, Academy of Uzbek. SSR, Tashkent, 1961 (In Russian).

39. Rao C. R. *Linear statistical inference and its applications.* Wiley, N.Y., 1965.

40. Rao C. R. and Toutenburg, F. *Linear Models*, Springer, N.Y., 1995.

41. Rao C. R. and Kleffe J. *Estimation of of variance components and its application.* North Holland Publ. Co., Amsterdam, 1988.

42. Shiryaev A. N. *Probability*, Springer-Verlag, Berlin, 1995.

43. Searle S. R. *Linear Models*, N.Y., Wiley, 1971.

44. Skene, A.M. Discrimination using latent structure models. *COMPSTAT 1978.* Physica-Verlag, Vienna, 199- 204, 1978.

45. Skene, A.M. (1980). Discussion of a paper by D.J.Bartolomew, *Journal of Royal Statistical Society* **B 42**, 314-315, 1980.

46. Whittle P. Bounds for the Moments of Linear and Quadratic Forms in Independent Variables, *Teor. Veroyatnost. i Primenen.*, **5(3)**, 331–335, 1960.

47. Wilks S. *Mathematical Statistics*, Wiley, 1962.

48. Wu J. Asymptotic Theory for Nonlinear Least Squares Estimation, *Ann. Math. Statist.*, **9**, 1981, 501-513.

49. Zhilinskaya E. I. Nonlinear LSE with predictors subject to errors, *Statistical Methods*, Perm, 1982, 19-48.

50. Zinn J. Note on Central Limit theorems in Banach spaces *Ann. Probab.*, bf 5, 1977, 283–286.