
DOI: https://doi.org/10.4213/tvp4537

Use of the all-Russian mathematical portal Math-Net.Ru implies that you have read and agreed to these terms of use
http://www.mathnet.ru/eng/agreement

Download details:
IP: 54.70.40.11
December 10, 2018, 07:26:58
ON THE SINGULARITIES OF THE INFORMATION MATRIX AND MULTIPATH CHANGE-POINT PROBLEMS

1. Introduction. The problem of singularities, that is, parameter values at which the information matrix is not of full rank, has been extensively studied by econometricians, electrical engineers, and, of course, statisticians. These studies can be divided into four categories: identification, inference at singularities, establishing nonsingularity of the information matrix in specific cases, and the investigation of the structure of the set of singularities.

1.1. Model identification. Studies on model identification date back to [42], [44], [66], [23], [26]. The seminal work of Rothenberg [57] put identification on a solid footing. He showed that local identifiability (i.e., identifiability on an open neighborhood of any given parameter value) is equivalent to
nonsingularity of the information matrix at the regular points of the parameter space (i.e., those points for which there is an open neighborhood such that the rank of the information matrix is constant). This result was further studied by Bowden [10]. White [68] and Sargan [60] provided detailed discussions on identifiability in econometrics. See also [27], [28]. Aldrich in [3] provides a thorough discussion on the pivotal role that identification plays in econometrics. He studies identification and some key statistical notions from both Bayesian and frequentist perspectives and presents their historical developments in statistics and econometrics. More recent accounts include [22] and [12], who are mainly concerned with nonparametric identification. See also [27], [28] for a conditional approach in identification, and [56] for an overview on econometrics and identification.

It transpires that there is an intimate link between identifiability and nonsingularity of the information matrix. This link has also been explored, amongst others, by Koopmans and Reiersol [43], Aitchison and Silvey [2], and its follow-up paper by Silvey [62], [61, p. 81] and Crowder [16, §4]. The identification and singularities of the information matrix have also been addressed in control theory by Stoica and Söderström [64], and more recently, by Stoica and Marzetta [63]. Stoica and Söderström [64] showed that there are cases where a singular information matrix does not contradict local identifiability.

The observation [64] seems, at first glance, in contradiction with [57]. However, Rothenberg [57] makes the crucial assumption of local constancy of the rank of the information matrix. Rothenberg’s result implies that singularity of the information matrix and local nonidentifiability can only occur if the rank of the information matrix is not locally constant. Our main results, Theorems 3 and 4, show that under identifiability and some tangible smoothness conditions, regularity is the rule rather than the exception. Rothenberg’s work may therefore be viewed as a precursor of the main results in this paper.

1.2. Inference at the singularities. Inference at the singularities is motivated mainly, but not exclusively, by statistical problems whose parameter space is a manifold with boundaries for which statistical inference is required on the boundary, where the information matrix is singular. A typical example of such cases is testing for the number of mixture components in finite mixture models. This challenging problem has generated interesting research in statistics. Dacunha-Castelle and Gassiat [18], [19] thoroughly studied this problem using conic parametrization. Likelihood ratio tests (LRT) and their asymptotic distributions for such problems have been extensively studied in the literature. We refer to [46], [47], and [29] for a recent account on LRT under loss of identifiability. Drton [20] introduces a unified approach to LRT and singularities through algebraic geometry. While the aforementioned references are mostly concerned with LRT, likelihood-
based asymptotic inference for a $p$-dimensional parameter, $\theta$, of a parametric model when the information matrix is of rank $p - 1$ at the true parameter value was taken by Rotnitzky, Cox, Bottai, and Robins [58]. They show, under usual smoothness conditions and identifiability, that $\sqrt{n}$-asymptotic inference is only available for $p - 1$ components of $\theta$, while the rate for the other component depends on the order of the first nonzero partial derivative of the log-likelihood with respect to that component. Klaassen and Lenstra [40], [41] show that the uniform $\sqrt{n}$-consistency neither locally nor globally exists in a semiparametric setting. Bottai [9] has studied confidence regions when the information matrix is singular.

1.3. Nonsingularity of the information matrix in specific cases. Nonsingularity of the information matrix is readily established for members of an exponential family provided that the sufficient statistics in the exponent are linearly independent. In many cases, however, nonsingularity does not immediately follow. Necessary and sufficient conditions for the nonsingularity of the information matrix in the case of several important models have been established. See [67] for an autoregressive model, [30] for three-layer perception networks, [51] for ARMA and ARIMA models, and [52] for binomial and Poisson mixed models.

1.4. Structure of the set of singularities. For many statistical models, nonsingularity of the information matrix seems virtually impossible to verify. An example of such models is a class of mixture models associated with multipath change-point problems (MCP), for which the form of the information matrix prevents the verification of this crucial condition. This is a general class of models which includes balanced longitudinal data and panel data as special cases. Despite its strong appeal, it seems that except for [17] and [40], not much research has been done on the structure of the set of singularities. When we are interested in making inference about the true parameter value which may or may not be a singularity, we may want to know how «reasonable» the assumption of nonsingularity is in the presence of other more tangible conditions. Theorem 3 shows that this condition is, in fact, «rarely» violated, provided that identifiability and some tangible smoothness conditions hold. The «rareness» should be interpreted in the intuitive geometric sense of perforation with holes. This «imprecise» statement is made explicit in Theorem 3.

Theorem 3 may be considered as the statistical counterpart of [11]. Next, one may proceed to study the Lebesgue measure of the set of singularities and establish the statistical counterpart of [59]. This result does not, however, hold under the assumptions of Theorem 3, as discussed by Klaassen and Lenstra [40]. The literature seems to incorrectly attribute such a result to Rothenberg [57] (see [58] and [14, p. 139]). [40] have indeed discussed the possibility of constructing, for any $\varepsilon > 0$, everywhere continuously differentiable and identifiable one-dimensional parametric models such that the set
on which the Fisher information is positive has measure $\leq \varepsilon$. Theorem 4 provides sufficient conditions for the set of singularities to be of measure zero.

1.5. Organization of the paper. The rest of this paper is organized as follows. Section 2 contains some classical results from differential topology and functional analysis which will be used in the sequel. The main result of Section 3, Theorem 3, shows that the set of singularities of the information matrix is a nowhere dense set. The Lebesgue measure of the set of singularities is treated in Section 4. Theorem 4 provides sufficient conditions for the set of singularities of the information matrix to be a nowhere dense set of measure zero. We study multipath change-point problems (MCP) in Section 5. Using the intrinsic Markovian structure of the change-point variable and the Cox proportional odds ratio model, we introduce a class of (semiparametric) models allowing the introduction of covariates. Theorems 5 and 6 of this section establish quasi-identifiability of this model, and hence pave the way towards the application of Theorems 3 and 4 to this model. Section 6 contains some concluding remarks.

2. Preliminaries. In this section we recall some standard results from analysis and geometry needed in the sequel. Theorem 1, the subimmersion theorem, and Theorem 2, the Upper semi-continuity of the spectrum, play key roles in the proof of Theorem 3 in Section 3. A proof of the subimmersion theorem can be found in [1, p. 205]. We refer the reader to [32, p. 167] for a proof of the upper semi-continuity of the spectrum. In the following theorem, $\mathcal{M}$ and $\mathcal{N}$ are Banach manifolds, $T_m f$ is the tangent of $f$ at $m$ and $\mathcal{R}(T_m f)$ is the kernel of $T_m f$.

**Theorem 1.** Suppose $f: \mathcal{M} \to \mathcal{N}$ is $C^r$ for $r \geq 1$, $n_0 \in \mathcal{N}$ and $f$ is a subimmersion in an open neighborhood of $f^{-1}(n_0)$. If $\mathcal{M}$ or $\mathcal{N}$ are finite dimensional, this is equivalent to $T_m f$ having constant rank in a neighborhood of each $m \in f^{-1}(n_0)$. Then $f^{-1}(n_0)$ is a submanifold of $\mathcal{M}$ with $T_m f^{-1}(n_0) = \mathcal{R}(T_m f)$.

The notion of subimmersion was introduced to formulate a geometric version of the implicit function theorem. Roughly speaking, a map $f: \mathcal{M} \to \mathcal{N}$ is a subimmersion if and only if $T_m f$, the tangent map of $f$, has locally constant rank. Thus instead of requiring that $f$ have a surjective tangent map at each $m \in \mathcal{M}$ in the implicit function theorem, we ask that the tangent map have locally constant rank. The subimmersion theorem then establishes existence of an open set in a subspace of dimension equal to the rank of the tangent map, on which one can find the «solution function» to an implicitly defined equation.

---

1) The subimmersion theorem is stated for $C^\infty$ in [1]. The result, however, holds for any $C^r$, $r \geq 1$. 

Definition 1. \( \mathcal{B} \) is called a Banach algebra if \( \mathcal{B} \) is an algebra as well as a Banach space and if, in addition, \( \| xy \| \leq \| x \| \| y \| \), for all \( x, y \in \mathcal{B} \).

Definition 2. Suppose \( \mathfrak{A} \) is an algebra with the unit element \( e \). An element \( x \) is called regular if there is an element \( x^{-1} \), called the inverse of \( x \), such that \( xx^{-1} = e \). A nonregular element is called singular.

Definition 3. Let \( \mathfrak{A} \) be a Banach space and let \( \mathcal{B} \) be the space of all bounded linear operators on \( \mathfrak{A} \), i.e., \( L(\mathfrak{A}, \mathfrak{A}) \), endowed with the operator norm topology. Suppose \( x \in \mathcal{B} \). The set of all values \( \lambda \) for which \( \lambda e - x \) is singular in the Banach algebra \( \mathcal{B} \) is called the spectrum of \( x \) and denoted by \( \sigma(x) \).

If \( \mathfrak{A} \) is a finite dimensional Banach space, then \( \sigma(x) \) is the set of the characteristic values of \( x \).

Definition 4. The spectrum of \( x \), \( \sigma(x) \), is called upper semicontinuous at \( x = a \) if for any open set \( O \) containing \( \sigma(a) \), there exists an \( \varepsilon > 0 \) such that \( \| x - a \| < \varepsilon \) implies that \( \sigma(x) \subset O \).

Theorem 2. Suppose \( \mathcal{B} \) is a Banach algebra and \( x \in \mathcal{B} \). Then the spectrum of \( x \) is an upper semicontinuous function of \( x \).

3. Singularities: a geometric perspective. In this section we study singularities of the information matrix from a geometric perspective. As is well known, the information matrix is the variance–covariance matrix of the partial derivatives of the log-likelihood function and hence is a positive semidefinite matrix. A positive semidefinite matrix is positive definite if and only if the determinant of the matrix is not zero. We, therefore, use positive definiteness and nonsingularity of the information matrix interchangeably in the sequel.

Theorem 3, the main result of this section, and its consequences, show that the set of singularities of the information matrix is a nowhere dense set. These results therefore imply that positive definiteness of the information matrix is «the rule» rather than «the exception» when the model is smooth and identifiable. One of our main motivations in this study is MCP and their associated mixture models. We will therefore pay special attention to finite mixture models in the sequel and discuss how the assumptions we make may be verified for such models.

Let \( \mathcal{P} = (\mathfrak{X}, \Theta, \mathfrak{B}(\mathfrak{X}), \mathfrak{B}(\Theta), P_\theta) \) be a statistical model, where \( \mathfrak{X} \) is the sample space and \( \Theta \) the parameter space, with corresponding \( \sigma \)-fields, \( \mathfrak{B}(\mathfrak{X}) \) and \( \mathfrak{B}(\Theta) \), respectively, and let \( P_\theta \) be a probability measure on \( \mathfrak{B}(\mathfrak{X}) \), for each \( \theta \in \Theta \).

A.1. The parameter space \( \Theta \subset \mathbb{R}^p \) is an open set.

A.2. The family \( \mathcal{P} = \{P_\theta; \theta \in \Theta\} \) is a family of probability measures on \((\mathfrak{X}, \mathfrak{B}(\mathfrak{X}))\) dominated by a \( \sigma \)-finite measure \( \mu \) and \( f(\cdot, \theta) \) is a Radon–Nykodim derivative of \( P_\theta \) with respect to \( \mu \).
The family $\mathcal{P}$ is identifiable with respect to $\theta$, i.e., $f(x, \theta) = f(x, \theta^*)$ a.e. $[\mu]$ implies that $\theta = \theta^*$.

Following [7] we embed $\mathcal{P}$ in $L_2(\mu)$ via the embedding $P_\theta \rightarrow s(\theta)$, where $s(\theta) = \sqrt{f(\cdot, \theta)}$.

**A.4.** The map $\theta \mapsto s(\theta)$ from $\Theta$ to $L_2(\mu)$ is Fréchet differentiable on $\Theta$, i.e., for any $\theta \in \Theta$ there exists a vector $Ds(\theta) = (Ds_1(\theta), \ldots, Ds_p(\theta))^T$ of elements of $L_2(\mu)$ such that

$$\|s(\theta + h) - s(\theta) - Ds^T(\theta) \cdot h\| = o(\|h\|) \quad \text{as} \quad \|h\| \rightarrow 0.$$ 

**A.5** The map $\theta \mapsto Ds_i(\theta)$ is a continuous map from $\Theta$ to $L_2(\mu)$ for $i = 1, 2, \ldots, p$.

Define the score function $\hat{\ell}(\theta)$ of an observation by

$$\hat{\ell}(\theta) = 2 \frac{Ds(\theta)}{s(\theta)} I_{\{s(\theta) > 0\}},$$

and the Fisher information matrix for $\theta$ by

$$I(\theta) = \int \hat{\ell}(\theta) \hat{\ell}^T(\theta) \, dP_\theta.$$ 

Let $\mathcal{M}_{p \times p}$ be the space of $p \times p$ real valued matrices endowed with the operator norm topology, where the operator norm is defined as follows. For any $A \in \mathcal{M}_{p \times p}$, define

$$\|A\|_{op} = \sup_{x \in \mathbb{R}^p \setminus \{0\}} \frac{\|Ax\|}{\|x\|}.$$ 

Note that $(\mathcal{M}_{p \times p}, \|\cdot\|_{op})$ with the usual addition and multiplication of matrices is a finite-dimensional Banach algebra.

Let $\phi(\cdot, \theta) = \ln f(\cdot, \theta)$ and define $\pi: \theta \mapsto \phi(\cdot, \theta)$. In the sequel, $D$ stands for Fréchet derivative, while $\nabla$ denotes the gradient.

**Lemma 1.** Suppose A.1–A.5 are fulfilled. Then

(i) $\int \hat{\ell}(\theta) \, dP_\theta = 0$, $\forall \theta \in \Theta$;

(ii) $I(\theta)$ is continuous on $\Theta$;

(iii) $D\pi(\theta) = \nabla \phi(x, \theta)$, a.e. $[\mu]$.

Proof. Part (i) follows from (9) of [7, p. 16]. Part (ii) follows from Proposition 3(A) of [7, p. 458] and the fact that $I$ is a map between two finite dimensional spaces. Part (iii) follows from Proposition 3(D) of [7, p. 458].

Lemma 1 is proved.

**Theorem 3.** Let $\Lambda = \{\theta \in \Theta: \det[I(\theta)] = 0\}$. Suppose A.1–A.5 hold. Then $\Lambda$ is a nowhere dense set.
Theorem 3 is proved in two steps. In the first step we show, using the continuity of \( I(\theta) \), that \( \Lambda \) is a closed set. We then show, in the second step, that \( \Lambda^\circ \), the interior of \( \Lambda \), is empty. The proof is based on a simple observation. Loosely speaking, if \( \text{rank}(D\pi) = \kappa < p \) on an open set, then \( p - \kappa \) variables are redundant. But this violates the identifiability of \( \mathcal{D} \), the family of the probability measures on \( (\mathcal{X}, \mathcal{B}(\mathcal{X})) \). Lemma 1 plays a key role in the proof of Theorem 3 below, since it allows us to obtain \( \text{rank}(D\pi) \) using the rank of the information matrix.

**Proof.** \( \text{Step 1.} \) \( \Lambda \) is a closed set. Let \( \theta \in \overline{\Lambda} \), the closure of \( \Lambda \). Then, there exists a sequence \( \{\theta_n\}_{n \in \mathbb{N}} \subset \Lambda \) such that \( \theta_n \to \theta \) as \( n \to \infty \). Thus, \( \det[I(\theta)] = \lim_{n \to \infty} \det[I(\theta_n)] = 0 \). Therefore, \( \det[I(\theta)] = 0 \) on \( \Lambda \) implies that \( \det[I(\theta)] = 0 \) on \( \overline{\Lambda} \).

**Step 2.** \( \Lambda^\circ = \emptyset \). Suppose \( \Lambda^\circ \neq \emptyset \). If rank \( I(\theta) \) = 0 on \( \Lambda \), then identifiability is violated, for all the partial derivatives must be zero on the open set \( \Lambda^\circ \). Therefore, suppose \( 0 < \text{rank}[I(\theta)] < p \) for some \( \theta \in \Lambda \). Let

\[
\kappa = \max\{l : \text{rank}[I(\theta)] = l, \text{ for some } \theta \in \Lambda\}.
\]  

Suppose \( \text{rank}[I(\theta_*)] = \kappa \) for some \( \theta_* \in \Lambda \). Using the fact that \( I(\theta) \) is a variance-covariance matrix of partial derivatives, there exist \( \kappa \) linearly independent partial derivatives at \( \theta_* \). Let \( \overline{I}(\theta_*) \) be the variance–covariance matrix of these \( \kappa \) linearly independent partial derivatives, and \( 0 < d_1(\theta_*) \leq \cdots \leq d_\kappa(\theta_*) \) be the eigenvalues of \( \overline{I}(\theta_*) \). Choose \( 0 < \zeta < d_1(\theta_*) \) and let \( \mathcal{D} = (d_1(\theta_*) - \zeta, d_\kappa(\theta_*) + \zeta) \). Thus \( \mathcal{D} \subset \mathbb{R} \) is an open set containing \( d_i(\theta_*) \), for \( i = 1, 2, \ldots, \kappa \). Using the upper semicontinuity of the spectrum (Theorem 2), there exists \( \varepsilon > 0 \) such that \( \|\overline{I}(\theta) - \overline{I}(\theta_*)\|_{\text{op}} < \varepsilon \) implies that the eigenvalues of \( \overline{I}(\theta) \) belong to \( \mathcal{D} \). Using the continuity of \( \overline{I} \), there exists an open set \( O_\varepsilon \subseteq \Lambda \) such that \( \|\overline{I}(\theta) - \overline{I}(\theta_*)\|_{\text{op}} < \varepsilon \), for all \( \theta \in O_\varepsilon \). Thus, \( \text{rank}[\overline{I}(\theta)] = \kappa \) for all \( \theta \in O_\varepsilon \). On the other hand, \( \kappa \) is the maximum rank of \( I \) on \( \Lambda \). Hence, \( \text{rank}[I(\theta)] = \kappa \) for all \( \theta \in O_\varepsilon \). Using the fact that there are only \( \kappa \) linearly independent partial derivatives on \( O_\varepsilon \), \( \text{rank}[T_{\theta_0}\pi] = \kappa \) for all \( \theta \in O_\varepsilon \). Now let \( n_0 = \ln f(x; \theta_0) \) for some \( \theta_0 \in O_\varepsilon \). By the «subimmersion theorem» (Theorem 1), \( T_{\theta_0}\pi^{-1}(n_0) = \mathcal{R}(T_{\theta_0}\pi) \). But identifiability implies that \( \pi \) is injective and so \( T_{\theta_0}\pi^{-1}(n_0) \) is of 0-dimension. On the other hand, \( \dim(\mathcal{R}(T_{\theta_0}\pi)) = p - \kappa \) and \( \kappa < p \). This is a contradiction. Hence \( \Lambda^\circ = \emptyset \). Theorem 3 is proved.

As discussed in the introduction, Rothenberg [57] established, under rather general conditions, a close tie between identifiability and positive definiteness of the information matrix at regular points of the parameter space. The crucial assumption of regularity, though seemingly natural, may be difficult to establish. Theorem 3 above addresses this issue and suggests that regularity is indeed «the rule» rather than «the exception» and, therefore, a natural assumption.
One can extend Theorem 3 to independent but not identically distributed random variables. Heterogeneity could, for example, be induced by the introduction of covariates in a model. Suppose that the covariate $Z$ is a random vector distributed according to $f_Z(z)$. Let $f_{X,Z}(x,z;\theta)$ be the joint distribution of $(X,Z)$ and $f_{X|Z}(x,\theta \mid z)$ the conditional density of $X$ given $Z = z$. A key condition for a direct extension is the identifiability of $f_{X|Z}(x;\theta \mid z)$ with respect to $\theta$. This assumption may fail to hold in many cases of interest and it is indeed restrictive in multipath change-point problems. An inspection of the proof of Theorem 3 shows that the identifiability assumption of the conditional distribution can be replaced by a weaker notion of identifiability, called quasi-identifiability, which often holds in MCP.

**Definition 5.** A collection of families of probability measures $\{\{P^z_\theta: \theta \in \Theta\}: z \in Z\}$ is called quasi-identifiable with respect to $\theta$ if for any $\theta, \theta^* \in \Theta$, $\theta \neq \theta^*$, it follows that there exists $\mathcal{Z}_{\theta,\theta^*} \in \mathcal{B}(Z)$, where $P(\mathcal{Z}_{\theta,\theta^*}) > 0$ such that $P^z_\theta \neq P^z_{\theta^*}$, for all $z \in \mathcal{Z}_{\theta,\theta^*}$.

Let $P^z_\theta(A) = \int_A f_{X|Z}(x;\theta \mid z)\mu(dx)$, for all $A \in \mathcal{B}(X)$ and let $\hat{\ell}_z(\theta)$ be the score function of $f_{X|Z}(x;\theta \mid z)$. Define $I_z(\theta) = \int \hat{\ell}_z(\theta)\hat{\ell}_z^T(\theta) dP^z_\theta$ and $I(\theta) = \mathbb{E}[I_Z(\theta)]$. The following corollary of Theorem 3 can readily be established.

**Corollary 1.** Let $Z$ be an ancillary for $\theta$. Suppose that $\{\{P^z_\theta: \theta \in \Theta\}: z \in Z\}$ is quasi-identifiable with respect to $\theta$ and for any $z \in Z$, Assumptions A.1–A.2 and A.4–A.5 are satisfied for $\{P^z_\theta: \theta \in \Theta\}$. Then the set of singularities of the information matrix $I(\theta)$ is a nowhere dense set.

**Remark.** As suggested by the referee, in view of the fact that quasi-identifiability is essentially equivalent to the identifiability of the joint distribution of $(X,Z)$, when $Z$ is ancillary for $\theta$, the above corollary follows immediately by taking the pair $(X,Z)$ to play the role of $X$. In many interesting applications, however, only the conditional model is known while the distribution of $Z$ is completely unknown. For this reason, it is useful to state the conditions in terms of the conditional model. Besides, the conditional approach seems more easily adaptable to the case of nonidentically distributed random variables when the heterogeneity is not necessarily induced by a covariate. For such extensions, one may consider $I(\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^n I_k(\theta)$, where $I_k(\theta)$ is the Fisher information contained in the $k$th sample. When the ancillarity assumption is violated, the joint and conditional approaches above are not equivalent. This assumption may be violated, for instance, in the presence of biased sampling. When sampling is biased and the marginal distribution of $Z$ is available, the joint distribution should be used. This issue has been recently discussed by McCullagh [49], and Bergerson, Asgharian, and Wolfson [6]. If, however, the marginal distribution of $Z$ is not available, the conditional approach seems to be appropriate, as dis-
4. Singularities: an analytic perspective. As was previously mentioned, Klaassen and Lenstra [40] have discussed the possibility of constructing everywhere continuously differentiable and identifiable one-dimensional parametric models for any $\varepsilon > 0$, such that the measure of the set of nonsingularities is $\leq \varepsilon$. We show, however, under more stringent assumptions the set of nonsingularities is a nowhere dense set of measure zero.

B.1. The parameter space $\Theta$ is a convex subset of $\mathbb{R}^p$.

B.2. Let $\xi = (\xi_1, \xi_2, \ldots, \xi_p)$. Suppose $f(x, \xi)$ and $\partial f(x, \xi) / \partial \xi_i$ are continuous functions on $\mathcal{X} \times \mathbb{R}$ for $i = 1, 2, \ldots, p$.

B.3 There are a function $k(x)$ and a constant $M$ for every compact set $K \subset \mathbb{R}^p$ such that

(i) $|\partial \phi_i(x, \xi) / \partial \xi_i|^2 \leq k(x)$ for $i = 1, 2, \ldots, p$,
(ii) $|\int_{\mathcal{X}} k(x) f(x, \xi) \nu(dx)| \leq M < \infty$ for all $\xi \in K$.

Theorem 4. Suppose $f(x, \xi) \neq 0$ on $\mathcal{X} \times \mathbb{R}$ is an analytic function of each of the $\xi_j = \theta_j + i\eta_j$, $j = 1, 2, \ldots, p$, for each $x \in \mathcal{X}$, in the region $R$ of parameter points for which $(\theta_1, \theta_2, \ldots, \theta_p) \in \Theta$. Suppose Assumptions A.1–A.3 and B.1–B.3 are fulfilled. Then $\Lambda$, the set of zeros of $\det[I(\theta)]$, is a nowhere dense set of Lebesgue measure zero.

Proof. First we notice that any convex set is path connected and therefore connected [21, p. 115, Theorem 5.3]. Using Hartogs’ theorem on analyticity in each variable [8, p. 140], it suffices to show that $E_\xi [B_{ij}(X, \xi)]$, where $B_{ij}(X, \xi) = \phi_i(X, \xi) \cdot \phi_j(X, \xi)$ and $\phi_i(x, \xi) = \partial \phi(X, \xi) / \partial \xi_i$, is separately analytic in each of the coordinates $\xi_1, \ldots, \xi_p$ for all $i, j = 1, 2, \ldots, p$. The Weierstrass convergence theorem now implies that $E_\xi [B_{ij}(X, \xi)]$ is an analytic function on $R$. Since $\det$ is an analytic function, it follows that $\det I(\xi)$ is an analytic map on $R$. On the other hand, Theorem 3 implies that $\det [I(\xi)]$ is a nonzero function on $R$. Thus $\Lambda$ is the set of zeros of a nonzero analytic function. This completes the proof of Theorem 4.

The literature on zero sets of analytic functions is vast. We know much more than what has been used in Theorem 4, i.e., being of measure zero. We refer to [13] for a brief account and to the classic book by Gunning and Rossi [31]. A more recent account can be found in [45].

The result of Theorem 4 is readily applicable to mixture distributions whose mixing distribution and components are analytic, provided that assumption B.3 holds. This condition may be cumbersome to verify, though it is still easier than directly checking nonsingularity of the information matrix for many mixture distributions. Working in the real domain, that is assuming that $\xi \in \mathbb{R}^p$ and $f(x, \xi)$ is analytic in $\xi$ for each $x \in \mathcal{X}$, one can reduce assumption B.3 to the same assumption imposed on mixing components and
Asgharian M.

mixing distributions, which will be much easier to verify. Real analyticity is, however, a more stringent assumption to impose on distributions.

5. Multipath change-point problems. In this section, we introduce the multipath change-point problem (MCP), a balanced longitudinal data setting incorporating subject-specific change-points. Using the intrinsic Markovian structure of a MCP, we show that a natural way to model a MCP is through finite mixture models. Having modeled the MCP, our task is to show that Theorems 3 and 4 are applicable to this setting, therefore paving the path for further statistical inference in the MCP. While the verification of Assumptions A.4 and A.5 for mixture models whose mixing distributions and components are smooth is not difficult (see [7, Proposition 2.1.1] and [33, Chap. 1, §7]), establishing identifiability (Assumption A.3) can be a challenge in many mixture models, and in particular, in the MCP and Mixture-of-Experts (MOE) models where mixing distributions (weights) may depend on covariates (see [34], [35], and [53]). We therefore confine our attention to establishing identifiability, and refer the reader to [4, Chap. IV, §5] for the verification of other conditions.

Subjects are typically monitored in follow-up studies for a period of time, resulting in a set of measurements indexed by time. Such data, broadly known as repeated measurements, have been extensively studied; see [15] and the references cited therein. Often a drug or stimulus is administered and the main concerns of the study are to make inference about the time when the drug takes effect, say the change-point, as well as the distribution before and after this change-point. If the effect of the drug were immediate, one could apply the standard methods of repeated measurements for statistical inference. In many cases, however, the change may be delayed for an unknown length of time or be gradual [37], [38].

In a single path change-point problem we have a sequence of random variables, \(X_1, X_2, \ldots, X_\tau, X_{\tau+1}, \ldots, X_m\) such that \(X_1, X_2, \ldots, X_\tau\) are realizations from a process, say \(\mathcal{P}_1\) whose joint distribution is \(F_1\) and that of \(X_{\tau+1}, X_{\tau+2}, \ldots, X_m\) is from \(\mathcal{P}_2\), with the joint distribution \(F_2\). If \(\tau\) is unknown and \(\tau < m\) we say that a change has occurred at \(\tau\), which is called the change-point.

In a multipath change-point problem, we have several paths, and associated with each path, a change-point. The observations in a multipath change-point setting form the matrix

\[
\begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1m} \\
X_{21} & X_{22} & \cdots & X_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nm}
\end{bmatrix}
\]

In the above matrix each row corresponds to one path. There are \(n\) paths
with \( m \) measurements on each. We say a change has taken place at \( \tau_i \), for the \( i \)th subject, when the first \( \tau_i \) measurements in the \( i \)th path, \( X_{i1}, X_{i2}, \ldots, X_{i\tau_i} \), come from a process, say \( \mathcal{P}_1 \), and the rest, \( X_{i\tau_i+1}, X_{i\tau_i+2}, \ldots, X_{im} \), from a different process, say \( \mathcal{P}_2 \). We always assume that \( \mathcal{P}_1 \) and \( \mathcal{P}_2 \) are independent. It should be emphasized that associated with each path we have a possibly different change-point. This is, of course, a rather general framework which covers panel data, balanced longitudinal data and repeated measurements as special cases, with the important twist that the time to effect of a stimulus might be different from one subject to another (see, e.g., [55]).

For a thorough discussion on the multipath change-point problem and its applications see [36]–[39] and [5].

In a typical MCP, we observe \((X, Z)\) on each subject, where \( X \) and \( Z \), respectively, represent the vector of measurements and the vector of covariates. Should we know where the change has occurred, i.e., conditioning on \( \tau = k \), the p.d.f. (p.m.f.) of \( X \) given \( Z = z \) and \( \tau = k \) is

\[
g_k^z(x; \gamma) = \left[ h_1^z(x_1; \gamma_{1,1}) \prod_{l=2}^{k} h_1^z(x_l; \gamma_{1,l|1,\ldots,l-1}) \right]
\times \left[ h_2^z(x_{k+1}; \gamma_{2,1}) \prod_{l=k+2}^{m} h_2^z(x_l; \gamma_{2,l|1,\ldots,l-1}) \right],
\]

for \( k = 1, 2, \ldots, m - 1 \) and

\[
g_m(x; \gamma) = h_1^z(x_1; \gamma_{1,1}) \prod_{l=2}^{m} h_1^z(x_l; \gamma_{1,l|1,\ldots,l-1}),
\]

where \( \gamma_{s,l|1,\ldots,l-1} \) for \( s = 1, 2 \) are, respectively, the subvectors of \( \gamma \) that parameterize the conditional distribution of \( X_l \) given \( X_1, \ldots, X_{l-1} \) before and after the change point. If the observations in each path are independent, then we have

\[
g_k^z(x, \gamma) = \prod_{l=1}^{k} h_1^z(x_l; \gamma_1) \prod_{l=k+1}^{m} h_2^z(x_l; \gamma_2), \quad k = 1, 2, \ldots, m - 1,
\]

and

\[
g_m^z(x, \gamma) = h_1^z(x_1; \gamma_1).
\]

The densities \( h_1^z \) and \( h_2^z \), respectively, represent the distribution of observations before and after the change point which we suppose are identifiable with \( \gamma_1 \) and \( \gamma_2 \), respectively. Thus, the full model given \( Z = z \) is

\[
f_z(x; \theta) = \sum_{k=1}^{m} \alpha_k(\beta; z) g_k^z(x; \gamma),
\]
where $\tau_i$ is the instant of change for the $i$th path, $\alpha_k(\beta, z) = P(\tau = k \mid Z = z)$, $g_k^*$ the joint distribution of the observations on the $i$th path given the covariates and $\tau = k$, $\theta = (\beta, \gamma)$, $\gamma = (\gamma_1, \gamma_2)$. The row vector $\beta$ is the vector of unknown regression parameters, while $\gamma_1$ and $\gamma_2$ are, respectively, the parameters of the distributions before and after the change point. For brevity we write $\alpha_k(\beta)$ or $\alpha_k$ for $\alpha_k(\beta, z)$, when there is no danger of confusion.

It should be noted that $\tau$, the change-point, is a latent variable and cannot be observed. Modeling $P_z(\tau = k) = P(\tau = k \mid Z = z)$ is, therefore, not a straightforward task. Having noticed the intrinsic Markovian structure of MCP, Asgharian and Wolfson [5] modeled $P(\tau = k \mid Z = z)$ as the hitting time distribution of a two-state Markov chain as follows,

$$
\alpha_k(\beta; z) = P_z(\tau = k; \beta) = \begin{cases} 
\pi_k(z) \prod_{l=0}^{k-1} [1 - \pi_l(z)] & \text{if } k = 1, 2, 3, \ldots, m - 1, \\
\prod_{l=0}^{m-1} [1 - \pi_l(z)] & \text{if } k = m,
\end{cases}
$$

(5.6)

where $\pi_k$ is the hazard of having a change-point at time $k$. We define $\pi_0 = 0$. In [5] several possibilities for modeling $\pi_k(z)$ were discussed and the constant logistic hazard model was analyzed, i.e.,

$$
\pi_k(z) = \pi(z) = \frac{\exp(\beta^Tz)}{1 + \exp(\beta^Tz)},
$$

(5.7)

in detail. Here we introduce a more flexible model and establish its quasi-identifiability, which is often the most subtle among the conditions of Theorems 3 and 4.

Inspired by the Cox proportional hazards model, one may introduce the following time dependent model:

$$
\pi_k(z) = \frac{\rho_k \exp(\beta^Tz)}{1 - \rho_k + \rho_k \exp(\beta^Tz)},
$$

(5.8)

which is essentially the discrete version of the Cox proportional hazards model. In model (5.8), $\rho_k$ should be interpreted as the base line hazard. Model (5.8) is, in fact, a proportional odds ratio model since

$$
\frac{\pi_k(z)}{1 - \pi_k(z)} = \frac{\rho_k}{1 - \rho_k} \exp(\beta^Tz).
$$

(5.9)

Model (5.5) also covers a class of models known as Mixture-of-Experts (MOE) models introduced in machine learning and neural networks. See [50, §5.13] and references cited therein for further discussions on MOE models.
Theorem 5. Let \( z \) be an \( r \)-vector of covariates, whose range contains 0 and at least one other vector. Suppose the design matrix \( M_s \) is of full rank and \( h_s^* \) for \( s = 1, 2 \) are, respectively, quasi-identifiable with respect to vector \( \gamma_s \) for \( s = 1, 2 \). Let

\[
\alpha_k(\beta, \rho; z) = \begin{cases} 
\frac{\rho_k \exp(\beta^T z)}{(1 - \rho_k) + \rho_k \exp(\beta^T z)} \sum_{j=0}^{k-1} \left[ 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right] & \text{if } k = 1, \ldots, m - 1, \\
\prod_{j=0}^{m-1} \left[ 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right] & \text{if } k = m, 
\end{cases}
\]

(5.10)

where \( \beta = (\beta_1, \ldots, \beta_r) \), \( \rho_0 = 0 \) and \( \rho = (\rho_1, \ldots, \rho_{m-1}) \) are the regression parameters. Then \( f_z(x; \theta) \) given by (5.3), (5.4), and (5.5) is quasi-identifiable with \( \theta = (\beta, \rho, \gamma) \).

Proof. Let \( \theta, \theta^* \in \Theta \) and \( \theta \neq \theta^* \). Suppose that \( f_z(x; \theta) = f_z(x; \theta^*) \) for all \( x \) and all \( z \). Following [5, Theorem 2] we obtain

\[
\frac{\alpha_1}{\alpha_1^*} = \frac{\alpha_2}{\alpha_2^*} = \cdots = \frac{\alpha_{m-1}}{\alpha_{m-1}^*}.
\]

(5.11)

For \( i = 1, 2, \ldots, m - 1 \), we have

\[
\frac{\alpha_i}{\alpha_i^*} = \frac{\rho_i \exp(\beta^T z)}{(1 - \rho_i) + \rho_i \exp(\beta^T z)} \prod_{j=0}^{i-1} \left( 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right) \frac{\rho_i^* \exp(\beta^* T z)}{(1 - \rho_i^*) + \rho_i^* \exp(\beta^* T z)} \prod_{j=0}^{i-1} \left( 1 - \frac{\rho_j^* \exp(\beta^* T z)}{(1 - \rho_j^*) + \rho_j^* \exp(\beta^* T z)} \right).
\]

Thus for \( i = 2, \ldots, m - 1 \) we have \( \alpha_i \rho_i^* = \alpha_i^* \rho_i^* \), which implies

\[
\frac{\rho_{i-1} \exp(\beta^T z)}{(1 - \rho_{i-1}) + \rho_{i-1} \exp(\beta^T z)} \prod_{j=0}^{i-2} \left( 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right) \frac{\rho_i^* \exp(\beta^* T z)}{(1 - \rho_i^*) + \rho_i^* \exp(\beta^* T z)} \prod_{j=0}^{i-1} \left( 1 - \frac{\rho_j^* \exp(\beta^* T z)}{(1 - \rho_j^*) + \rho_j^* \exp(\beta^* T z)} \right) = \frac{\rho_{i-1} \exp(\beta^T z)}{(1 - \rho_{i-1}) + \rho_{i-1} \exp(\beta^T z)} \prod_{j=0}^{i-1} \left( 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right) \frac{\rho_i^* \exp(\beta^* T z)}{(1 - \rho_i^*) + \rho_i^* \exp(\beta^* T z)} \prod_{j=0}^{i-2} \left( 1 - \frac{\rho_j^* \exp(\beta^* T z)}{(1 - \rho_j^*) + \rho_j^* \exp(\beta^* T z)} \right) \prod_{j=0}^{i} \left( 1 - \frac{\rho_j \exp(\beta^T z)}{(1 - \rho_j) + \rho_j \exp(\beta^T z)} \right) \prod_{j=0}^{i-1} \left( 1 - \frac{\rho_j^* \exp(\beta^* T z)}{(1 - \rho_j^*) + \rho_j^* \exp(\beta^* T z)} \right)
\]

Simplifying the above expressions we obtain

\[
\frac{\rho_{i-1} \rho_i^*}{(1 - \rho_i^*) + \rho_i^* \exp(\beta^* T z)} = \frac{\rho_{i-1} \rho_i}{(1 - \rho_i) + \rho_i \exp(\beta^T z)}.
\]
That is,
\[
(\rho_{i-1}^\ast)^{-1}(1 - \rho_{i-1}^\ast) \quad \frac{(\rho_i^\ast)^{-1}(1 - \rho_i^\ast) + \exp(\beta^T z)}{\rho_i^{-1}(1 - \rho_i) + \exp(\beta^T z)},
\]
for \( i = 2, 3, \ldots, m - 1 \). Define
\[
\psi_i = \text{odds}(\rho_i) = \frac{\rho_i}{1 - \rho_i}.
\]
Then using (5.12), we have
\[
\exp(\beta^T z) = \frac{(\psi_{i-1}^\ast)^{-1}[\psi_i^{-1} + \exp(\beta^T z)]}{\psi_i^{-1}} - (\psi_i^\ast)^{-1},
\]
for \( i = 2, 3, \ldots, m - 1 \). Suppose \( z_1 \) and \( z_2 \) are two distinct vectors in the range of \( z \) such that \( \beta^T (z_1 - z_2) \neq 0 \). Thus, using (5.13) we have
\[
\frac{\exp(\beta^T z_1) - \exp(\beta^T z_2)}{\exp(\beta^T z_1) - \exp(\beta^T z_2)} = \frac{(\psi_i^\ast)^{-1}}{\psi_i^{-1}}, \quad \text{for } i = 1, \ldots, m - 2.
\]
Letting \( z = 0 \) in (5.13), we obtain
\[
\frac{1 + (\psi_i^\ast)^{-1}}{1 + \psi_i^{-1}} = \frac{(\psi_{i-1}^\ast)^{-1}}{\psi_{i-1}^{-1}}, \quad \text{for } i = 2, \ldots, m - 1.
\]
It then follows from (5.14) and (5.15) that
\[
\frac{1 + (\psi_i^\ast)^{-1}}{1 + \psi_i^{-1}} = \frac{(\psi_i^\ast)^{-1}}{\psi_i^{-1}}, \quad \text{for } i = 2, \ldots, m - 2,
\]
which implies that \( \psi_i = \psi_i^\ast \) for \( i = 2, \ldots, m - 2 \). On the other hand, letting \( z_1 = 0 \) in (5.14), we have for any \( \beta^T z \neq 0 \)
\[
\frac{1 - \exp(\beta^T z)}{1 - \exp(\beta^T z)} = \frac{(\psi_i^\ast)^{-1}}{\psi_i^{-1}} = 1, \quad \text{for } i = 1, \ldots, m - 2,
\]
and hence \( M_\beta(\beta - \beta^\ast)^T = 0 \), which implies \( \beta = \beta^\ast \). It then remains to show that \( \psi_{m-1} = \psi_{m-1}^\ast \). This latter step follows from (5.11) and the form of \( \alpha_k(\beta, z) \) by noticing that the ratios in (5.11) should all be equal to 1 now. One needs only consider that \( \alpha_{m-2}/\alpha_{m-2}^\ast = \alpha_{m-1}/\alpha_{m-1}^\ast \). Identifiability of \( \gamma \), parameters involved in before and after the change distributions, follows from
the same arguments as in [5]. Thus θ = θ∗. This is a contradiction and the proof of Theorem 5 is complete.

Within each path independence, relations (5.3) and (5.4), assumed in Theorem 5, while tenable in some applications, is prone to failure in some others. We relax this assumption in Theorem 6. Suppose that the observations in each path are realizations of a Markov process. Such a framework in the single path setting has been discussed by Yakir [69] for finite state space Markov chains (see also [65]). The basic assumption is that the processes before and after the change point are independent with different transition probabilities. In our notation, this means

\[ g^*_k(x; γ) = \left[ h^*_1(x_1; λ_1) \prod_{l=2}^{k} h^*_1(x_l; λ_l, ρ_l \mid x_{l-1}) \right] \times \left[ h^*_2(x_{k+1}; λ_2) \prod_{l=k+2}^{m} h^*_2(x_l; λ_l, ρ_l \mid x_{l-1}) \right], \]

(5.17)

for \( k = 1, 2, \ldots, m - 1 \), and

\[ g^*_m(x; γ) = h^*_1(x_1; λ_1) \prod_{l=1}^{m} h^*_1(x_l; λ_l, ρ_l \mid x_{l-1}), \]

(5.18)

where \( γ_s = (λ_s, ρ_s) \) for \( s = 1, 2 \). The components of \( γ_s \), i.e., \( λ_s \) and \( ρ_s \), are, respectively, the unknown parameters of the marginal and transition probability functions. Theorem 6 extends Theorem 5 for stationary Markov processes provided that \( λ_s \) and \( ρ_s \) are, respectively, identifiable from \( h^*_s(x; λ_s) \) and \( h^*_s(x; γ_s \mid y) \).

**Theorem 6.** Suppose that the measurements before and after the change, respectively, follow independent stationary Markov processes \( P_1 \) and \( P_2 \) whose marginal and transition probability functions are quasi-identifiable. Then \( f_z(x; θ) \) given by (5.5), (5.17), and (5.18), is quasi-identifiable with respect to \( θ \).

**Proof.** Suppose \( f_z(x; θ) = f_z(x; θ^*) \) for almost all \( x \). As in the proof of Theorem 5 we only work with marginal distributions. Since \( P_s \) for \( s = 1, 2 \) are stationary processes with quasi-identifiable marginal and transition probability functions, using a similar argument as that used in the proof of Theorem 5, we obtain \( λ_s = λ^*_s \), for \( s = 1, 2 \) and \( β = β^* \).

To complete the proof we must show that \( ρ_s = ρ^*_s \) for \( s = 1, 2 \). It suffices to apply the same technique as in [36] and [5] and integrate \( x_3, \ldots, x_m \) out. One then obtains

\[
\begin{align*}
\alpha_1 h_1^*(x_1; λ_1) h_2^*(x_2; λ_2) &+ (1 - \alpha_1) h_1^*(x_1; λ_1) h_1^*(x_2; λ_1, ρ_1 \mid x_1) \\
= \alpha_1 h_1^*(x_1; λ_1^*) h_2^*(x_2; λ_2^*) &+ (1 - \alpha_1^*) h_1^*(x_1; λ_1^*) h_1^*(x_2; λ_1^*, ρ_1^* \mid x_1). \quad (5.19)
\end{align*}
\]
We have already shown that $\lambda_s = \lambda'_s$ for $s = 1, 2$ and $\beta = \beta'$. Therefore, (5.19) is reduced to $h^*_i(x_2; \lambda_1, \rho_1 \mid x_1) = h^*_i(x_2; \lambda^*_i, \rho^*_i \mid x_1)$. Thus, the identifiability of the transition probability function implies that $\rho_1 = \rho^*_1$. Similarly, by integrating over $x_1, \ldots, x_{m-2}$, we obtain

$$(1 - \alpha_m - \alpha^*_{m-1})h^*_2(x_{m-1}; \lambda_2)h^*_2(x_m; \lambda_2, \rho_2 \mid x_{m-1})$$

$$+ \alpha^*_{m-1}h^*_2(x_{m-1}; \lambda_2^*)h^*_2(x_m; \lambda_2^*, \rho_2^* \mid x_{m-1})$$

$$= (1 - \alpha^*_m - \alpha^*_{m-1})h^*_2(x_{m-1}; \lambda_2^*)h^*_2(x_m; \lambda_2^*, \rho_2^* \mid x_{m-1})$$

$$+ \alpha^*_m h^*_1(x_{m-1}; \lambda_1^*)h^*_2(x_m; \lambda_2^*) + \alpha^*_m h^*_1(x_{m-1}; \lambda_1^*)h^*_1(x_m; \lambda_1^*, \rho_1^* \mid x_{m-1}).$$

(5.20)

Using (5.20) and the identifiability of the transition probability function of $\mathcal{P}_2$ implies $\rho_2 = \rho_2^*$. This completes the proof of Theorem 6.

Example 1. Suppose the observations before and after the change are generated by stationary Markov processes $\mathcal{P}_s^{z, \eta}$, $s = 1, 2$, defined for the $i$-th subject by

$$X_{s,i}^z(t) = \mu_s + \eta_{s,i} + \zeta_s z_i + Y_s(t),$$

(5.21)

where $z$ is a continuous and $\eta$ is a qualitative covariate. The process $Y_s(t)$ is an AR(1) process defined by

$$Y_s(t) = \rho_s Y_s(t - 1) + \varepsilon_s(t),$$

(5.22)

where $0 < |\rho_s| < 1$ and $\{\varepsilon_s(t)\}_t$ is a sequence of independent and identically distributed random variables distributed according to $N(0, \sigma^2_s)$ for $s = 1, 2$. Define $\lambda_s = (\mu_s, \eta_s, \zeta_s, \sigma_s)$, for $s = 1, 2$. It is not difficult to see that all the conditions of Theorem 6 are fulfilled and therefore we have quasi-identifiability of the mixture model given by (5.5) with respect to $\theta = (\rho_1, \lambda_1, \rho_2, \lambda_2)$.

6. Closing remarks. 1. Oxtoby and Ulam [54] have shown that sets of the first Baire category (i.e., meager sets) and sets of measure zero in $[0, 1]^r$ are equivalent, in the sense that the set of automorphisms that take a set of the first Baire category to a set of measure zero is a residual subset of all automorphisms on $[0, 1]^r$. This result, in conjunction with Klaassen and Lenstra [40] and our Theorem 3 imply that we can introduce two different parametrizations such that under one $\Lambda$ is a nowhere dense set, but of positive measure, while under the other one it is a nowhere dense set of measure zero. We note, however, as also mentioned by Oxtoby and Ulam [54], such maps cannot be $C^1$, though the maps can be almost everywhere differentiable.

2. To prove Theorem 3 we assumed that $\Theta \subset \mathbb{R}^p$ is a bounded open set. This assumption can be replaced by $\Theta$ is a smooth manifold of dimension $p$. Indeed, we need only assume a $C^2$ differentiable structure on
Singularities of the information matrix and MCP

Θ. An inspection of the proof of Theorem 3 also shows that the identifiability assumption can be replaced by \( T_\theta, \pi^{-1}(n_0) \) is of 0-dimension, where \( n_0 = \ln f(x, \theta_0) \), for some \( \theta_0 \in \Theta \). In light of this observation, we can conclude that \( \Lambda \) is a nowhere dense set for finite mixture distributions that are identifiable up to permutations.

Acknowledgments. The author would like to express his sincere gratitude to an anonymous referee whose comments and suggestions considerably improved this manuscript. He would also like to thank his colleagues David Wolfson and Abbas Khalili as well as his former students Dr. Marco Carone, Didier Chetelat, and Jun Shi for their comments on earlier drafts of this manuscript.

REFERENCES


Поступила в редакцию 28.IV.2009
Исправленный вариант 6.II.2012