

The asymptotic effect of tuning parameters - Supplementary material

Ingrid Dæhlen^{1,2}, Nils Lid Hjort¹ and Ingrid Hobæk Haff¹

¹ Department of Mathematics, University of Oslo

² Norwegian Computing Center, Post box 114 Blindern, Oslo, 0314, Norway

Address for correspondence: Ingrid Dæhlen, Department of Mathematics, University of Oslo, Moltke Moes vei 35, 0851 Oslo, Norway, email: ingrdae@math.uio.no

S1 Proof of Lemma 1 and Theorem 6

We start by showing a lemma regarding uniform consistency.

Lemma S1.1. *Let $Y \in \mathbb{R}^d$, $\theta \in \mathbb{R}^p$ and $\lambda \in \Lambda \subseteq \mathbb{R}^q$ where Λ is a compact set. With $\Psi_{\lambda,n}: \mathbb{R}^{d+p} \rightarrow \mathbb{R}^p$ and $\Psi_\lambda: \mathbb{R}^p \rightarrow \mathbb{R}^p$ for each $\lambda \in \Lambda$, let $\theta_0(\lambda)$ be the unique zero of $\Psi_\lambda(\theta)$ for each fixed λ and assume the following.*

(C1) *There exists $\epsilon > 0$ such that $\Psi_{\lambda,n} \xrightarrow{\text{Pr}} \Psi_\lambda$ uniformly on $\mathcal{S}_\epsilon = \{(\theta, \lambda) \mid \lambda \in \Lambda, \|\theta_0(\lambda) - \theta\| \leq \epsilon\}$*

(C2) *The function $(\theta, \lambda) \mapsto \Psi_\lambda(\theta)$ is continuously differentiable on $\mathcal{S}_\epsilon$ with $\Psi'_\lambda\{\theta(\lambda)\}$ non-singular for each $\lambda \in \Lambda$.*

Then, if $\Psi_{\lambda,n}(\theta)$ has at most a single zero $\hat{\theta}(\lambda)$ for each n and each λ ,

$$\sup_{\lambda} \|\hat{\theta}(\lambda) - \theta_0(\lambda)\| \xrightarrow{\text{Pr}} 0. \quad (\text{S1.1})$$

Proof. The following is a modified version of the proof of theorem 5.42 in Van der Vaart (2000).

By (C2) and the implicit function theorem, $\theta(\lambda)$ is continuous as a function of λ in Λ . Hence, $\{\theta(\lambda) \mid \lambda \in \Lambda\}$ is a compact set. This implies that $\mathcal{S}_\epsilon$ is compact. Because of this, the extreme value theorem and the fact that $\Psi'_\lambda\{\theta(\lambda)\}$ is non-singular for all

$\lambda \in \Lambda$, we can choose ϵ to be small enough that $\Psi'_\lambda(\theta)$ is non-singular for all $(\theta, \lambda) \in \mathcal{S}_\epsilon$, ensuring that $(\theta, \lambda) \mapsto \Psi'_\lambda(\theta)^{-1}$ is continuous on $\mathcal{S}_\epsilon$. By the extreme value theorem, we then have that the norm of each component of $\Psi'_\lambda(\theta)^{-1}$ is bounded by some $C < \infty$ in $\mathcal{S}_\epsilon$.

Now fix some $\delta > 0$ and choose δ_1 smaller than both ϵ and $\delta/2C$. Arguing as in Van der Vaart (2000), we see that there exists closed neighbourhoods $G_{\delta_1}^\lambda$ of $\theta_0(\lambda)$ such that Ψ^λ is a homeomorphism from $G_{\delta_1}^\lambda$ into the closed ball around zero with radius δ_1 : $\overline{B_{\delta_1}(0)}$ for each λ . By construction of δ_1 and the inverse function theorem, the partial derivatives of the inverse of Ψ_λ is bounded by C for each λ . Hence, the diameter of $G_{\delta_1}^\lambda$ is bounded by $2C\delta_1 < \delta$ for each λ implying $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ . We will show that $\Psi_{\lambda,n}(\theta)$ has a root in $G_{\delta_1}^\lambda$ for every λ with probability tending to one. Since $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ , this would guarantee (S1.1).

Since $\delta_1 < \epsilon$, we know

$$\sup_{\lambda \in \Lambda} \sup_{\theta \in G_{\delta_1}^\lambda} \|\Psi_{\lambda,n}(\theta) - \Psi_\lambda(\theta)\| \xrightarrow{\text{Pr}} 0.$$

Let $K_{\delta_1,n}$ be the event that $\|\Psi_{\lambda,n}(\theta) - \Psi_\lambda(\theta)\| < \delta_1$ for all $\theta \in G_{\delta_1}^\lambda$ and $\lambda \in \Lambda$. Then the above implies that $\Pr(K_{\delta_1,n}) \rightarrow 1$.

Now let $y \in \overline{B_{\delta_1}(0)}$. Since Ψ_λ is a homeomorphism from $G_{\delta_1}^\lambda$ into $\overline{B_{\delta_1}(0)}$ for each λ , $\Psi_\lambda^{-1}(y) \in G_{\delta_1}^\lambda$ for every $\lambda \in \Lambda$. Hence, on the event $K_{\delta_1,n}$, we have $\|\Psi_{\lambda,n}\{\Psi_\lambda^{-1}(y)\} - y\| \leq \delta_1$ for all $\lambda \in \Lambda$, which implies that $y \mapsto y - \Psi_{\lambda,n} \circ \Psi_\lambda^{-1}(y)$ maps $\overline{B_{\delta_1}(0)}$ into itself for every $\lambda \in \Lambda$. Since the maps are continuous, Brouwer's fixed point theorem ensures that $y \mapsto y - \Psi_{\lambda,n} \circ \Psi_\lambda^{-1}(y)$ has a fixed point in $\overline{B_{\delta_1}(0)}$ for every $\lambda \in \Lambda$. This is equivalent to $\Psi_{\lambda,n}$ having a zero in $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ . Because of this, the probability of all $\Psi_{n,\lambda}$ having a zero within δ of $\theta_0(\lambda)$ is bounded by $\Pr(K_{\delta_1,n})$. As this latter probability tends to one by the previous arguments, (S1.1) follows. This concludes the proof. ■

When $\Psi_\lambda(\theta) = \mathbb{E} \phi(Z, \theta, \lambda)$ and $\Psi_{\lambda,n}(\theta) = n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function

$\phi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$, condition (C1) can be easily verified using the uniform law of large numbers. We state this in the following corollary.

Corollary S1.2. *Let $\Psi_\lambda(\theta) = \mathbb{E} \phi(Z, \theta, \lambda)$ and $\Psi_{\lambda,n}(\theta, \lambda) = n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function $\phi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$. Then (C1) follows from*

(C3) *There exists an F -integrable function $m_0: \mathbb{R}^d \rightarrow \mathbb{R}$ and an $\epsilon > 0$ such that $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$ for all $(\theta, \lambda) \in S_\epsilon$ and F -almost all Z .*

We will prove a theorem characterizing the distance from $\hat{\theta}(\lambda)$ to $\hat{\theta}_{(-i)}(\lambda)$.

Lemma S1.3. *For each λ in Λ , a compact subset of $\mathbb{R}^q$ and $1 \leq i \leq n$ let $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$ be the solutions to $\sum_{j=1}^n \phi(Z_j, \theta, \lambda)$ and $\sum_{j \neq i} \phi(Z_j, \theta, \lambda)$ in Θ , a compact subset of $\mathbb{R}^p$. Assume there exist functions m_0 and m_1 such that $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$ and $\|\partial_\theta \phi\{z, \theta, \lambda\}\| \leq m_1(z)$ for all θ, λ and almost all z . Then, if $\theta \mapsto \mathbb{E} \phi(Z, \theta, \lambda)$ satisfies (C2) of Lemma S1.1 and the eigenvalues of $J(\theta, \lambda)^{-1} = -\{\mathbb{E} \partial_\theta \phi(Z, \theta, \lambda)\}^{-1}$ are bounded in $\Theta \times \Lambda$, the following holds for all λ :*

$$\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq a_{0,n} m_0(Z_i) \quad (\text{S1.2})$$

where $a_{0,n} = O_p(1/n)$ and does not depend on i or λ . If, furthermore, the eigenvalues of $\partial_\theta J(\theta, \lambda)$ are bounded in $\Theta \times \Lambda$, we have

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = n^{-1} J\{\theta_0(\lambda), \lambda\}^{-1} \phi\{Z_i, \theta_0(\lambda), \lambda\} + E_n^i(\lambda) \quad (\text{S1.3})$$

where $E_n^i(\lambda)$ satisfies

$$\|E_n^i(\lambda)\| \leq n^{-1} \{a_{1,n} + a_{2,n} m_0(Z_i) + a_{3,n} m_0(Z_i)^2 + a_{4,n} m_0(Z_i)^2 m_1(Z_i)\} \quad (\text{S1.4})$$

with $a_{j,n} = o_p(1)$ uniformly in λ , not depending on i for $j = 1, 2, 3, 4$.

Proof. By definition we have

$$0 = \sum_{j=1}^n \phi\{Z_j, \hat{\theta}(\lambda), \lambda\} \quad \text{and} \quad 0 = \sum_{j \neq i}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\}. \quad (\text{S1.5})$$

Hence, a Taylor expansion of $\sum_{j=1}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\}$ around $\hat{\theta}(\lambda)$ reveals

$$\begin{aligned} \phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} &= \sum_{j=1}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\} \\ &= \sum_{j=1}^n \phi\{Z_j, \hat{\theta}(\lambda), \lambda\} + \left(\sum_{j=1}^n \partial_{\theta_i^*} \phi\{Z_j, \theta, \lambda\} \right) \{\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\} \\ &= -n J_n\{\theta_i^*(\lambda), \lambda\} \{\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\} \end{aligned}$$

for some $\theta_i^*(\lambda)$ s on the line segment between $\hat{\theta}_{(-i)}(\lambda)$ and $\hat{\theta}(\lambda)$, and with $J_n\{\theta, \lambda\} = -n^{-1} \sum_{j=1}^n \partial_{\theta} \phi(Z_j, \theta, \lambda)$. Because of the above, we can write

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1} J_n\{\theta_i^*(\lambda), \lambda\}^{-1} \phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}.$$

Since there exists a function $m_1(z)$ bounding all partial derivatives of $\phi(z, \theta, \lambda)$ with respect to θ , $J_n(\theta, \lambda)$ converges uniformly to $J(\theta, \lambda) = -E \partial_{\theta} \phi(Z, \theta, \lambda)$ over the compact parameter space $\Theta \times \Lambda$. Furthermore matrix inversion is uniformly continuous over compact sets not containing non-invertible matrices, ensuring that also $-J_n(\theta, \lambda)^{-1}$ converges to $-J(\theta, \lambda)^{-1} = E \partial_{\theta} \phi(Z, \theta, \lambda)^{-1}$ uniformly in λ and θ . Then $-J_n\{\theta_i^*(\lambda), \lambda\}^{-1} = -J\{\theta_i^*(\lambda), \lambda\}^{-1} + \epsilon_n^i$ where $\epsilon_n^i = o_p(1)$ uniformly in λ and i . By assumption, the eigenvalues of $J(\theta, \lambda)^{-1}$ are bounded by a number with absolute value C_1 . Then $\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq n^{-1}(C_1 + \delta_n) \|\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}\|$, where $\delta_n = o_p(1)$ does not depend on λ or i . Lastly by assumption $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$. The above then guarantees

$$\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq n^{-1}(C_1 + \delta_n) m_0(Z_i) \quad (\text{S1.6})$$

for all λ . This proves (S1.2).

By the previous arguments

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1}[J\{\theta_i^*(\lambda), \lambda\}^{-1} + \epsilon_n^i]\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}. \quad (\text{S1.7})$$

where $\epsilon_n^i = o_p(1)$ uniformly in i and λ . We will again use Taylor expansions to arrive at a result. We start with the second factor on the right hand side: $J\{\theta_i^*(\lambda), \lambda\}^{-1} = J\{\theta_0(\lambda), \lambda\}^{-1} + (\partial_{\theta_i^{**}(\lambda)} J\{\theta, \lambda\}^{-1})\{\theta_i^*(\lambda) - \theta_0(\lambda)\}$, where $\theta_i^{**}(\lambda)$ is some vector on the line segment between $\theta_0(\lambda)$ and $\theta_i^*(\lambda)$. Now since $\theta_i^*(\lambda)$ lies on the line segment between $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$, $\|\theta_i^*(\lambda) - \theta_0(\lambda)\| \leq \|\hat{\theta}(\lambda) - \theta_0(\lambda)\| + \|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\|$. Combining this with Lemma S1.1 and (S1.6) ensures $\|\theta_i^*(\lambda) - \theta_0(\lambda)\| \leq R_n + n^{-1}(C_1 + \delta_n)m_0(Z_i)$, where $R_n = o_p(1)$ uniformly in λ and does not depend on i . In addition, the eigenvalues of $\partial_\theta J(\theta, \lambda)^{-1}$ are bounded by C_2 . Hence, $J\{\theta_i^*(\lambda), \lambda\}^{-1} = J\{\theta_0(\lambda), \lambda\}^{-1} + r_n^i(\lambda)$ where $\|r_n^i(\lambda)\| \leq n^{-1}(C_1 + \delta_n)C_2m_0(Z_i)$ uniformly in λ .

Secondly, we will work with the third factor on the right hand side of (S1.7). Taylor expanding the function $\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}$ around $\theta_0(\lambda)$, reveals

$$\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} = \phi\{Z_i, \theta_0(\lambda), \lambda\} + \partial_{\theta_i^{***}}\phi(Z_i, \theta, \lambda)\{\hat{\theta}_{(-i)}(\lambda) - \theta_0(\lambda)\},$$

for some $\theta_i^{***}(\lambda)$ on the line segment between $\theta_{(-i)}(\lambda)$ and $\theta_0(\lambda)$. Since there exists $m_1(z)$ bounding the norm of each component of $\partial_\theta \phi(z, \theta, \lambda)$ and depending only on z , we can use arguments similar to those above to show $\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} = \phi\{Z_i, \theta_0(\lambda), \lambda\} + s_n^i(\lambda)$ where $\|s_n^i(\lambda)\| \leq n^{-1}(C_1 + \delta_n)m_1(Z_i)m_0(Z_i)$ uniformly in λ and for all i .

We are now ready to complete the proof. Combining all of the previous arguments shows

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1}[J\{\theta_0(\lambda), \lambda\}^{-1} + r_n^i(\lambda) + \epsilon_n^i][\phi\{Z_i, \theta_0(\lambda), \lambda\} + s_n^i(\lambda)],$$

which imply $\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1}J\{\theta_0(\lambda), \lambda\}\phi\{Z_i, \theta_0(\lambda), \lambda\} + E_n^i(\lambda)$ where

$$\|E_n^i(\lambda)\| \leq n^{-1}\{a_{1,n} + a_{2,n}m_0(Z_i) + a_{3,n}m_0(Z_i)^2 + a_{4,n}m_0(Z_i)^2m_1(Z_i)\}$$

and $a_{j,n} = o_p(1)$ uniformly in λ and does not depend on i for $j = 1, 2, 3, 4$. This concludes the proof. ■

The first part of Lemma S1.3 shows Lemma 1 in the article. We will now use the second part to show Theorem 6.

S1.1 Proof of Theorem 6

Taylor's theorem ensures, $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \sum_{i=1}^n \psi(Z_i, \hat{\theta}) + \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T (\hat{\theta}_{(-i)} - \hat{\theta})$ for some θ_i^* s on the line segments between $\hat{\theta}$ and $\hat{\theta}_{(-i)}$. By assumption, all conditions of Lemma S1.3 hold true, and hence,

$$\sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T (\hat{\theta}_{(-i)} - \hat{\theta}) = -n^{-1} \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T J^{-1} \varphi(Z_i, \theta_0) + \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T E_N^i. \quad (\text{S1.8})$$

By the same lemma and existence of p_1 , we have

$$\begin{aligned} \left\| \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T E_N^i \right\| &\leq a_{1,n} n^{-1} \sum_{i=1}^n p_1(Z_i) + a_{2,n} n^{-1} \sum_{i=1}^n m_0(Z_i) p_1(Z_i) \\ &\quad + a_{3,n} n^{-1} \sum_{i=1}^n m_0(Z_i)^2 p_1(Z_i) + a_{4,n} n^{-1} \sum_{i=1}^n m_0(Z_i)^2 m_1(Z_i) p_1(Z_i). \end{aligned}$$

Since the moments of all terms in the above equation exists by assumption and $a_{j,n} = o_p(1)$ for $j = 1, 2, 3, 4$, the second term on the right hand side of (S1.8) is $o_p(1)$.

For the first term of (S1.8), note that by Taylor's theorem $\nabla_{\theta_i^*} \psi(Z_i, \theta) = \nabla_{\theta_0} \psi(Z_i, \theta) + H_{\theta_i^{**}} \psi(Z_i, \theta) (\theta_i^* - \theta_0)$ for some θ_i^{**} on the line segment between θ_i^* and θ_0 . By the triangle inequality and the fact that θ_i^* lies on the line segment between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$,

$\|\theta_i^* - \theta_0\| \leq \|\hat{\theta} - \theta_0\| + \|\hat{\theta} - \hat{\theta}_{(-i)}\|$. Hence, by Lemma S1.3, we have

$$\left\| -n^{-1} \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T J(\theta_0)^{-1} \varphi(Z_i, \theta_0) - n^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta)^T J(\theta_0)^{-1} \varphi(Z_i, \theta_0) \right\| \leq \|\hat{\theta} - \theta_0\| n^{-1} C \sum_{i=1}^n p_2(Z_i) m_0(Z_i) + a_{0,n} n^{-1} C \sum_{i=1}^n p_2(Z_i) m_0(Z_i)^2,$$

for some C . This expression is $o_p(1)$ since the expected value of both means are finite and $\|\hat{\theta} - \theta_0\|, a_{0,n} = o_p(1)$ by Lemma S1.3. This concludes the proof.

S2 Appendix - When the limit function is flat

In Theorem 1, we assumed that λ_0 was the unique minimizer of the limiting risk $R(\lambda) = E \psi\{Z, \theta_0(\lambda)\}$. In certain cases, however, $\theta_0(\lambda)$, the solution to $E \varphi\{Z, \theta, \lambda\} = 0$, is constant as a function of λ , and as a result, $R(\lambda)$ is flat. This is the case for hybrid models on the form $\varphi(z, \theta, \lambda) = \lambda \varphi_1(z, \theta) + (1 - \lambda) \varphi_2(z, \theta)$ when $E \varphi_1(Z, \theta)$ and $E \varphi_2(Z, \theta)$ share roots, e.g. in the hybrid generative-discriminative model when the generative model is specified correctly or for hybrid combinations of empirical and hybrid likelihood functions (see Dæhlen and Hjort (2025)) under model conditions. In fact, when $R(\lambda)$ is flat, the limit of both $TE(\lambda)$ and $CV(\lambda)$ are constant, and there is no obvious “best” value for λ , making it unclear what $\hat{\lambda}$ even should be aiming for were it consistent. We will now use informal mathematical arguments and simulations to illustrate why a clear and informative limit result for $\hat{\theta}(\hat{\lambda})$ is hard to find when $\theta_0(\lambda)$ is constant.

Assume that $\theta_0(\lambda)$ is constant and equal to θ_0 . To investigate the behaviour of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ we will first take a closer look at the function $CV(\lambda)$. By Theorem 6, the following holds pointwise,

$$CV(\lambda) = n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda), \lambda\} - n^{-1} \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0, \lambda)\} + o_p(1/n).$$

Taylor expanding the first term on the right hand side in the above equation, reveals

$$\begin{aligned} n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda), \lambda\} &= n^{-1} \sum_{i=1}^n \psi(Z_i, \theta_0, \lambda) + \{\hat{\theta}(\lambda) - \theta_0\}^T n^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta, \lambda) \\ &\quad + \{\hat{\theta}(\lambda) - \theta_0\}^T \frac{1}{2n} \sum_{i=1}^n H_{\theta^*} \psi(Z_i, \theta, \lambda) \{\hat{\theta}(\lambda) - \theta_0\}, \end{aligned}$$

for some θ^* on the line segment between $\hat{\theta}(\lambda)$ and θ_0 . Assume θ_0 is the true minimizer of $E \psi(Z, \theta)$, i.e. that $E \nabla_{\theta_0} \psi(Z, \theta) = 0$. Then, arguments similar to those given in the proof of Lemma 4.2 and Theorem 6 show

$$\begin{aligned} S_n(\lambda) &= n \left(\text{CV}(\lambda) - \frac{1}{n} \sum_{i=1}^n \psi(Z_i, \theta_0) \right) \\ &= U_n(\lambda)^T V_n - \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0)\} + \frac{1}{2} U_n(\lambda)^T W_n U_n(\lambda) + o_p(1), \end{aligned}$$

where $U_n(\lambda) = \sqrt{n}\{\hat{\theta}(\lambda) - \theta_0\}$, $V_n = n^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta)$, and $W_n = n^{-1} \sum_{i=1}^n H_{\theta_0} \psi(Z_i, \theta)$. Since θ_0 is constant with respect to λ , we have $\hat{\lambda} = \text{argmin}_{\lambda} S_n(\lambda)$.

Note that W_n is consistent for $W = E H_{\theta_0} \psi(Z, \theta)$ by the assumptions of Theorem 6. Furthermore, it does not depend on λ . If, in addition, the conditions of Van der Vaart (2000, Ex. 19.7) are fulfilled for $(z, \lambda) \mapsto (J(\theta_0, \lambda)^{-1} \varphi(z, \theta_0, \lambda), \psi(z, \theta_0))$, the process $\lambda \mapsto (U_n(\lambda), V_n)$ converges as a process to a Gaussian process $(U(\lambda), V)$ with covariance structure as described in the beginning of Van der Vaart (2000, Sec. 19.2). This, in combination with Slutsky's lemma and the continuous mapping theorem, is sufficient for S_n to converge in distribution in Skorokhod space to the following limit process

$$S(\lambda) = U(\lambda)^T V - T(\lambda) + \frac{1}{2} U(\lambda)^T W U(\lambda), \quad (\text{S2.9})$$

where $T(\lambda) = \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0)\}$.

Direct computations show $E S(\lambda) = (1/2) \text{Tr}\{W J(\lambda)^{-1} K(\lambda) J(\lambda)^{-1}\}$. Hence, if there exists a λ_0 such that $\hat{\theta}(\lambda_0)$ is equal to the maximum likelihood estimator, the

Cramer-Rao theorem guarantees that λ_0 minimizes $ES(\lambda)$. Since such a λ_0 often exists in the case where $\theta_0(\lambda)$ is constant, it is tempting to think that $\hat{\lambda}$ converges in probability to this λ_0 . Sadly this conclusion cannot be drawn, as S is a random process and there is no guarantee that minimizing the expected value of S minimizes the process itself. Furthermore, the convergence of S_n towards S is in distribution. Because of this, convergence in probability of $\hat{\lambda}$ towards λ_L , the minimizer of S , does not follow. We can only deduce $\hat{\lambda} \xrightarrow{d} \lambda_L$, and since S is a stochastic process and not a nonrandom function, its minimizer is not guaranteed to be a constant. Because of this, λ_L is a random variable and $\hat{\lambda} \xrightarrow{d} \lambda_L$ does not imply convergence in probability. The above therefore does not show consistency of $\hat{\lambda}$ towards λ_0 or any fixed value for that matter, but rather that the distribution of $\hat{\lambda}$ tends towards that of the minimizer of S .

To illustrate the above, we performed a small simulation experiment for a hybrid LDA-logistic regression model with one covariate and binary response variable. We simulated from an LDA model, under which the true parameter minimizes both the limit of the generative and logistic likelihood. Hence $\theta_0(\lambda)$ is constantly equal to the true parameter in this simulation. We used a $\psi(y, x, \theta) = (y - \exp\{\beta(\theta)^T x\} / [1 + \exp\{\beta(\theta)^T x\}])^2$, a smoothed version of the 0 - 1 loss function. With $n = 2000$ we computed $B = 1000$ different realizations of S_n . A selection of sample paths is displayed in Figure 1 together with a histogram of the corresponding minima. From the figure we see that the sample paths are indeed not constant and that although the minimum is most often one, other values are repeatedly chosen, implying that λ_L might really not be constant in this case.

Although the above arguments and simulations seem to indicate that there is reasonable doubt that $\hat{\lambda}$ is consistent for any one value, this does not necessarily rule out all hope that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ has a normal limiting distribution. In the simulation experiment described above, we also computed $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ in each iteration. Histograms of each component of this vector are displayed in Figure 2. From the plots we note that the distributions at least look quite normal. This intuition

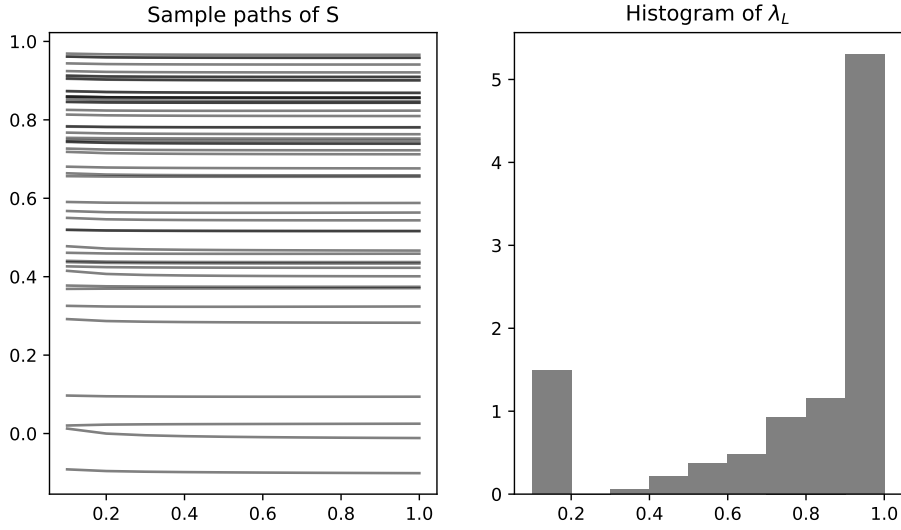


Figure 1: The figure shows two plots. To the left $B = 1000$ realizations of S_n with $n = 2000$ are displayed. On the right hand side a histogram of the corresponding minima is shown.

is supported by the QQ-plots shown in Figure 3. Looking at these plots, we see no obvious deviances from the corresponding normal distributions.

One possible solution is to work with $A_n(\hat{\lambda}) = \sqrt{n}K(\theta_0, \lambda_0)^{-1/2}J(\theta_0, \hat{\lambda})\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ rather than $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$. Pointwise, we have $A_n(\lambda) \xrightarrow{d} N(0, I_p)$, and hence, we would expect $A_n(\hat{\lambda}) \xrightarrow{d} N(0, I_p)$ as well. To check whether this idea seems reasonable, we computed $A_n(\hat{\lambda})$ in each iteration of the simulation setting described above. A histogram of each component of $A_n(\hat{\lambda})$ is shown in Figure 4 together with the graph of density function for the standard normal distribution. The plot seem to indicate that $A_n(\hat{\lambda})$ behaves more or less like a standard normally distributed variable.

One could argue that Corollary 1 and Theorem 5 failing to cover situations where the limiting risk function is flat, makes the result and the estimators of Theorem 4 less useful in practical situations. Say that we for instance, have fitted a generative hybrid-discriminative model to some data and want to apply the theory in the main part of this paper to construct confidence intervals and hypothesis tests for the parameters in the model. If the generative model fitted to the data is wrongly specified, this can be

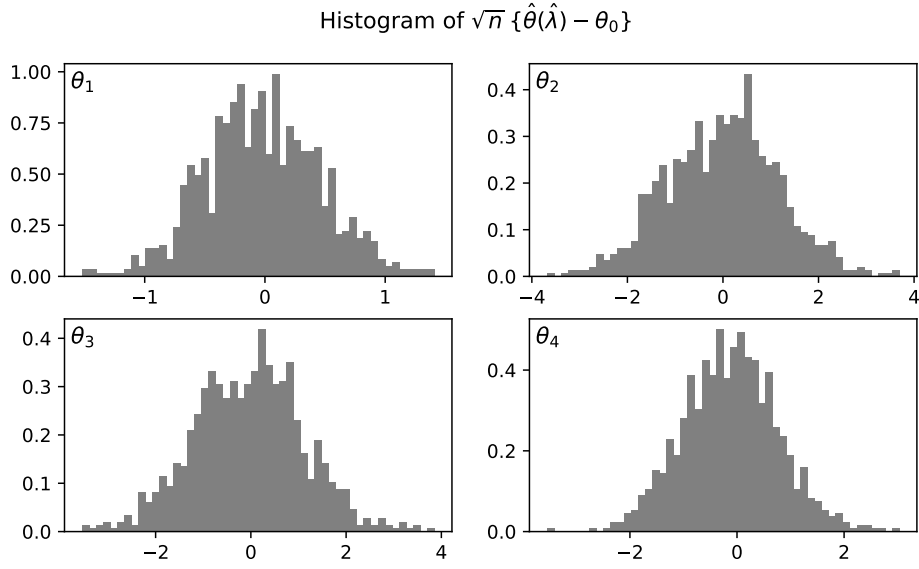


Figure 2: The figure shows histograms of each component of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

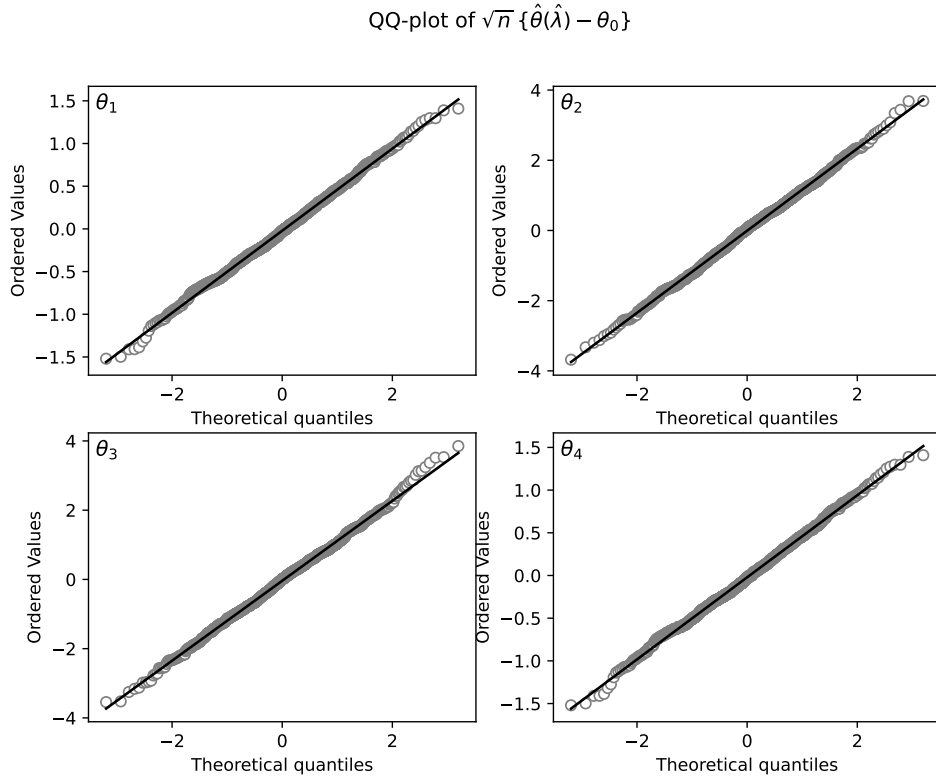


Figure 3: The figure shows QQ-plots for each component of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$. The empirical quantiles are compared to those of fitted normal distributions.

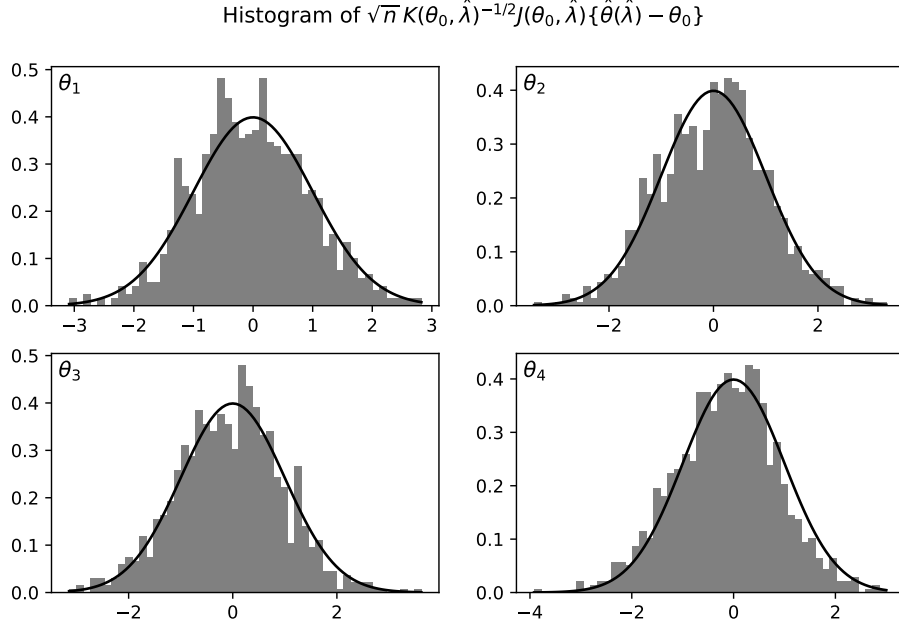


Figure 4: The figure shows histograms of each component of $\sqrt{n} K(\theta_0, \hat{\lambda})^{-1/2} J(\theta_0, \hat{\lambda}) \{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

done by applying Corollary 1 or Theorem 5 in conjunction with Theorem 4. If the model conditions do hold true, however, the arguments made in this section shows that the results cannot be applied. Hence, we cannot say that Corollary 1, Theorem 5 or the estimators in Theorem 4 are truly model-agnostic, since they fail to apply when the limiting risk function is flat. In theory, this is of course quite problematic. In practice, however, this is less of a problem. When $\theta'_0(\lambda) = 0$ for all values of λ , $E\varphi(Z, \theta_0, \lambda) = 0$ for all λ . Because of this, $\hat{\theta}'(\hat{\lambda})$ of Theorem 3 is consistent for 0 and $\hat{A}^*$ in Theorem 4 is asymptotically equivalent with $(J^{-1}(\hat{\lambda}), 0, 0)$. Hence, $\hat{A}^* \hat{K}(\hat{\lambda})^* (\hat{A}^*)^T$ is asymptotically equivalent with $J^{-1}(\hat{\lambda}) K(\hat{\lambda}) J^{-1}(\hat{\lambda})$ which is what the simulations of this section seem to indicate is the limiting variance of $\sqrt{n} \{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

S3 The effect of validation sets

In this section we will give an informal discussion on how an alternative tuning scheme based on setting aside validation sets affects the asymptotic distribution of $\hat{\theta}(\hat{\lambda})$. This

problem is essentially equivalent to situations with auxiliary information, see Qin (2000) which have inspired our proofs.

Assume we have i.i.d. data $Z_1, \dots, Z_n$ from some distribution F . In this section, we will assume that the first n_1 data points are used for tuning λ and that the remaining $n_2 = n - n_1$ data points are used for estimating $\hat{\theta}$ with the chosen value of λ , i.e.

$$\hat{\theta}(\lambda) = \text{zero of } \sum_{i=n_1+1}^n \varphi(Z_i, \theta, \lambda) \text{ and } \hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{n_1} \psi[Z_i, \hat{\theta}(\lambda)].$$

Arguing as in the paragraph preceding Theorem 1 in the main part of the paper, we get

$$\hat{\alpha} = \begin{pmatrix} \hat{\theta}(\hat{\lambda}) \\ \hat{\lambda} \\ \hat{\theta}'(\hat{\lambda}) \end{pmatrix} = \text{zero of } \begin{pmatrix} \sum_{i=n_1+1}^n \eta_1(Z_i, \theta, \lambda, D) \\ \sum_{i=1}^{n_1} \eta_2(Z_i, \theta, \lambda, D) \\ \sum_{i=n_1+1}^n \eta_3(Z_i, \theta, \lambda, D) \end{pmatrix},$$

where η_1 , η_2 and η_3 are as defined in Theorem 1 in the main part of the paper.

By Lemma S1.1, $\hat{\alpha}$ converges in probability towards α_0 , the zero of

$$\alpha \mapsto \begin{pmatrix} p \operatorname{E} \eta_1(Z, \alpha) \\ (1-p) \operatorname{E} \eta_2(Z, \alpha) \\ p \operatorname{E} \eta_3(Z, \alpha) \end{pmatrix},$$

provided $n_1/n \rightarrow p$ as $n \rightarrow \infty$, and the regularity conditions of Lemma S1.1 hold true. Since p is constant, α_0 is equal to the zero of $\alpha \mapsto \operatorname{E}[(\eta_1(Z, \alpha)^T, \eta_2(Z, \alpha)^T, \eta_3(Z, \alpha)^T)]^T$. Hence, $\hat{\alpha}$, and in particular $\hat{\theta}(\hat{\lambda})$, is consistent also when tuning with respect to the error on a validation set.

For asymptotic normality, notice that Taylor's theorem ensures

$$0 = \begin{pmatrix} \sum_{i=n_1+1}^n \eta_1(Z_i, \hat{\alpha}) \\ \sum_{i=1}^{n_1} \eta_2(Z_i, \hat{\alpha}) \\ \sum_{i=n_1+1}^n \eta_3(Z_i, \hat{\alpha}) \end{pmatrix} = \begin{pmatrix} \sum_{i=n_1+1}^n \eta_1(Z_i, \alpha_0) \\ \sum_{i=1}^{n_1} \eta_2(Z_i, \alpha_0) \\ \sum_{i=n_1+1}^n \eta_3(Z_i, \alpha_0) \end{pmatrix} + (\hat{\alpha} - \alpha_0)^T \begin{pmatrix} \sum_{i=n_1+1}^n J\eta_1(Z_i, \alpha^*) \\ \sum_{i=1}^{n_1} J\eta_2(Z_i, \alpha^*) \\ \sum_{i=n_1+1}^n J\eta_3(Z_i, \alpha^*) \end{pmatrix},$$

for some α^* on the line segment between $\hat{\alpha}$ and α_0 . Hence,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = - \left[\frac{1}{n} \begin{pmatrix} \sum_{i=n_1+1}^n J\eta_1(Z_i, \alpha^*) \\ \sum_{i=1}^{n_1} J\eta_2(Z_i, \alpha^*) \\ \sum_{i=n_1+1}^n J\eta_3(Z_i, \alpha^*) \end{pmatrix}^{-1} \right]^T \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{i=n_1+1}^n \eta_1(Z_i, \alpha_0) \\ \sum_{i=1}^{n_1} \eta_2(Z_i, \alpha_0) \\ \sum_{i=n_1+1}^n \eta_3(Z_i, \alpha_0) \end{pmatrix}.$$

Assuming $J\eta_j$ for $j = 1, 2, 3$ are sufficiently regular, we therefore have

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = - \left[\begin{pmatrix} p E J\eta_1(Z, \alpha_0) \\ (1-p) E J\eta_2(Z, \alpha_0) \\ p E J\eta_3(Z, \alpha_0) \end{pmatrix}^{-1} \right]^T \begin{pmatrix} p^{1/2} n_2^{-1/2} \sum_{i=n_1+1}^n \eta_1(Z_i, \alpha_0) \\ (1-p)^{1/2} n_1^{-1/2} \sum_{i=1}^{n_1} \eta_2(Z_i, \alpha_0) \\ p^{1/2} n_2^{-1/2} \sum_{i=n_1+1}^n \eta_3(Z_i, \alpha_0) \end{pmatrix} + o_p(1),$$

which by the CLT converges to a central multivariate normal distribution with covariance matrix $-[J\Psi^p(\alpha_0)^{-1}]^T \text{Var} \eta^p(Z, \alpha_0) [J\Psi^p(\alpha_0)^{-1}]^T$, where

$$\begin{aligned} \eta^p(Z, \alpha) &= (p^{1/2} \eta_1(Z, \alpha)^T, (1-p)^{1/2} \eta_2(Z, \alpha)^T, p^{1/2} \eta_3(Z, \alpha)^T)^T \text{ and} \\ \Psi^p(\alpha) &= E(p\eta_1(Z, \alpha)^T, (1-p)\eta_2(Z, \alpha)^T, p\eta_3(Z, \alpha)^T)^T. \end{aligned}$$

Now let $P = \text{diag}\{p, (1-p), p\}$ since $\eta_p(z, \alpha) = P^{1/2} \eta(z, \alpha)$ and $\Psi^p(\alpha) = P\Psi(\alpha)$, the limiting variance of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is

$$- [J\Psi(\alpha_0)^{-1}]^T P^{-1/2} \text{Var} \eta(Z, \alpha_0) P^{-1/2} [J\Psi(\alpha_0)^{-1}]^T, \quad (\text{S3.10})$$

where Ψ and η are as defined in Theorem 1 in the main part of the article, compared to

$$- [J\Psi(\alpha_0)^{-1}]^T \text{Var} \eta(Z, \alpha_0) [J\Psi(\alpha_0)^{-1}]^T,$$

when cross-validation or training error is used for tuning.

Assuming $0 < p < 1$, the entries of $P^{-1/2}$ are all greater than one, which makes the variance of $\hat{\alpha}$ greater when using validation sets rather than cross-validation or training

error to tune λ . This is a relatively natural result as setting aside a validation set will use the data less effectively, which affects the speed of convergence accordingly. As a reduction in convergence speed from $\sqrt{n}$ to $\sqrt{pn}$ is equivalent to a relative variance change of p , (S3.10) reflects this reduced convergence speed.

References

- Dæhlen, I. and N. L. Hjort (2025). Model robust hybrid likelihood. *Journal of Statistical Planning and Inference* (241), 106327.
- Qin, J. (2000). Miscellanea. combining parametric and empirical likelihoods. *Biometrika* 87(2), 484–490.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.