

1 Proofs for Section 2

In this section we will give a formal proof of Theorem 2.1. The arguments are similar to those given by e.g. Imbens et al. (1998) and Schennach (2007), and start by finding asymptotic results for the vector $(\lambda_n(\hat{\mu}), \hat{\mu})$. This is achieved by utilizing techniques of proofs for estimators fitted with two-stage procedures, see e.g. Joe (2005).

First, however, we state and prove a result concerning consistency of Z-estimators. The result is a slight modification of Theorem 5.42 in Van der Vaart (2000).

Lemma 1.1. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and let $\psi: \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^q$ be a function with $\Theta \subseteq \mathbb{R}^p$. Assume there is a unique solution θ_n to the equation $0 = \Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Y_i, \theta)$ for all n and a unique solution θ_0 to the limit equation $0 = \Psi(\theta) = E\psi(Y, \theta)$ when $Y \sim F$. Then θ_n converges in probability to θ_0 , provided*

(C1) the function Ψ is continuously differentiable in a neighborhood of θ_0 and $\Psi'(\theta_0)$ is invertible;

(C2) there exists a neighborhood $\mathcal{N}$ of θ_0 and an F -integrable function $p_0: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\|\psi(y, \theta)\| \leq p_0(y)$ for all $\theta \in \mathcal{N}$ and F almost all y .

Proof. Let Ψ and Ψ_n be defined as in the theorem. Close inspection of the proof of Theorem 5.42 in Van der Vaart (2000) shows that the conditions of the theorem are slightly too strict and that the only thing needed for the arguments to go through is that Ψ is continuously differentiable with non-singular derivative in a neighborhood of θ_0 and that there exists a closed ball around θ_0 inside which Ψ_n converges uniformly to Ψ . The first assessment is ensured by (C1) and continuity of the determinant function. The second condition follows from (C2) and the uniform law of large numbers. Hence, there exists a sequence of estimators converging in probability to θ_0 . By assumption, $\Psi_n(\theta) = 0$ has a

unique solution θ_n . Because of this, θ_n must be this consistent sequence, proving that θ_n converges in probability to θ_0 . $\square$

Before we prove Theorem 2.1, we will demonstrate an initial related result.

Theorem 1.2. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ be a function continuously differentiable in the last s arguments. Let $\hat{\mu}$ be the unique maximizer of the empirical likelihood function, constructed using m and the data $Y_1, \dots, Y_n$, $\hat{\lambda}$ the unique solution to*

$$0 = n^{-1} \sum_{i=1}^n \frac{m(Y_i, \hat{\mu})}{1 + \lambda^T m(Y_i, \hat{\mu})},$$

and $\hat{\alpha} = (\hat{\lambda}, \hat{\mu})$. Lastly, set $\psi_{\text{EL}}(y, \alpha) = [1 + \lambda^T m(y, \mu)]^{-1} (m(y, \mu)^T, \lambda^T \partial m(y, \mu))^T$ with ∂m as short hand for $\partial m / \partial \mu$. Then $\hat{\alpha}$ is consistent for $\alpha_0 = (\lambda_0, \mu_0)$, the unique solution to $0 = \Gamma^{\text{EL}}(\lambda, \mu) = \mathbb{E} \psi_{\text{EL}}(Y, \alpha)$, where $Y \sim F$. Furthermore,

$$\hat{\alpha} = \alpha_0 - n^{-1} J \Gamma^{\text{EL}}(\alpha_0)^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (1)$$

provided the following conditions hold true:

- (1) *the function Γ^{EL} is continuously differentiable in a neighborhood of α_0 and its Jacobian at α_0 is non-singular;*
- (2) *there exists $L > 0$ and a neighborhood $\mathcal{N}$ of α_0 on which $\Pr(1 + \lambda^T m(Y, \mu) > L) = 1$ for all $(\lambda, \mu) \in \mathcal{N}$;*
- (3) *there exists a neighborhood $\mathcal{N}$ of α_0 and functions p_j for $j = 0, 1, 2$ such that for F -almost all y and all $\alpha \in \mathcal{N}$, the norm of m and all second and third partial order derivatives of m are bounded by p_j for $j = 0, 1, 2$ respectively, with $\mathbb{E} p_0(Y)^4, \mathbb{E} p_1(Y)^4 < \infty$ and $\mathbb{E} p_2(y)^2 < \infty$.*

In particular, $\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N[0, J \Gamma^{\text{EL}}(\alpha_0)^{-1} K_{\text{EL}} J \Gamma^{\text{EL}}(\alpha_0)^{-1}]$, with $K_{\text{EL}} = \text{Var} \psi_{\text{EL}}(Y, \alpha_0)$, $Y \sim F$.

Proof. If m is differentiable, the MEL estimator satisfies the following equation

$$0 = \frac{\partial}{\partial \mu} \log \text{EL}_n(\mu) = - \sum_{i=1}^n \frac{\lambda'_n(\mu)^T m(Y_i, \mu) + \partial m(Y_i, \mu)^T \lambda_n(\mu)}{1 + \lambda_n(\mu)^T m(Y_i, \mu)} = - \sum_{i=1}^n \frac{\partial m(Y_i, \mu)^T \lambda_n(\mu)}{1 + \lambda_n(\mu)^T m(Y_i, \mu)},$$

where the last equality follows from definition of $\lambda_n(\mu)$. Hence, $(\lambda_n(\hat{\mu}), \hat{\mu})$ is the solution to $0 = n^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha)$, where ψ_{EL} is as defined in the theorem. This shows that $\hat{\alpha}$ is a Z-estimator, and hence, standard results concerning such estimators can be used to study limiting properties of $\hat{\alpha}$.

We start by showing consistency towards α_0 . This is achieved by proving condition (C2) of Lemma 1.1. (C1) holds by condition (1) of this theorem. Fix a neighborhood $\mathcal{N}$ of α_0 such that all conditions of the theorem hold true on $\mathcal{N}$. By assumption (2) there exists $L > 0$ such that $1 + \lambda^T M(\mu) > L$ with probability 1 for all $\alpha \in \mathcal{N}$. Combining this with condition (3) ensures $\|\psi_{\text{EL}}(y, \alpha)\| \leq C_1 p_0(y) + C_2 p_1(y)$ for some constants $C_1, C_2 \in \mathbb{R}$. The right hand side of this inequality is a sum of F -integrable functions. This shows condition (C2) of Lemma 1.1, and guarantees that $\hat{\alpha}$ converges in probability to α_0 .

We will use Theorem 5.21 in Van der Vaart (2000) to show (1). We start with the Lipschitz condition. Direct computations show $\|\frac{\partial \psi}{\partial \alpha}(y, \alpha)\| \leq C_3 p_0(y)^2 + C_4 p_0(y) p_1(y) + C_5 p_1(y) + C_6 p_1(y)^2 + C_7 p_2(y)$, for some constants C_j for $j = 3, \dots, 7$. This is a consequence of condition (3) of the theorem. Since all functions on the right hand side of this equation are square integrable with respect to the measure F , the sum is square integrable as well. This in conjunction with the mean value theorem for multivariate functions, ensures the Lipschitz property of Theorem 5.21 in Van der Vaart (2000). Furthermore, $E\|\psi(Y, \alpha_0)\|^2 < \infty$ by a similar argument. The rest of the conditions of Theorem 5.21 in Van der Vaart (2000) hold by assumption. (1) follows.

Strictly speaking, we have only showed that the sequence of solutions to $n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha)$ satisfies the properties of the theorem. Since, $\log \text{EL}_n(\mu)$ is equal to $-\sum_{i=1}^n \log[1 + \lambda_n(\mu)^T m(Y_i, \mu)]$ only when $1 + \lambda_n(\mu)^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$, $\partial \log \text{EL}_n(\hat{\mu}) / \partial \mu = 0$ is not immediately equivalent to $n^{-1} \sum_{i=1}^n \psi(Y_i, \hat{\alpha}) = 0$. Since $1 + \lambda^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$ holds on $\mathcal{N}$ by assumption (2), however, $\hat{\alpha}$ will eventually satisfy $1 + \lambda_n(\mu)^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$ and $\log \text{EL}_n(\mu)$ will be equal to $\sum_{i=1}^n \log[1 + \lambda_n(\mu)^T m(Y_i, \mu)]$ with probability tending to one. As a result the root of $n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha)$ will eventually be equal to $\hat{\alpha}$. $\square$

1.1 Proof of Theorem 2.1

By Theorem 1.2, $\hat{\mu}$ converges in probability to the last s components of α_0 , the parameter solving $0 = \text{E}\psi_{\text{EL}}(Y, \alpha)$. By definition $\lambda(\mu)$ is the root of $\text{E}m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]$, the first q component functions in $\text{E}\psi_{\text{EL}}(Y, \alpha)$. Hence, $\alpha_0 = (\lambda(\mu_0), \mu_0)$ where μ_0 solves $0 = \text{E}\partial m(Y, \mu)^T \lambda(\mu)/[1 + \lambda(\mu)^T m(Y, \mu)]$. Furthermore, the maximizer of $-\text{E}\log[1 + \lambda(\mu)^T m(Y, \mu)]$ satisfies $0 = \text{E}[\lambda'(\mu)^T m(Y, \mu) + \partial m(Y, \mu)^T \lambda(\mu)]/[1 + \lambda(\mu)^T m(Y, \mu)]$. Since $\lambda(\mu)$ solves $0 = \text{E}\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$, the right hand side of this equation is equal to $\text{E}\partial m(Y, \mu)^T \lambda(\mu)/[1 + \lambda(\mu)^T m(Y, \mu)]$ which, by definition of μ_0 , has μ_0 as a root. In particular, μ_0 maximizes $-\text{E}\log[1 + \lambda(\mu)^T m(Y, \mu)]$.

For (6) in the main part of the article, let $(0, I)$ be the $(q + s) \times (q + s)$ block matrix where I is the $s \times s$ -identity matrix. Then, by (1),

$$\hat{\mu} = (0, I)\hat{\alpha} = \mu_0 - (0, I)\Gamma^{\text{EL}}(\alpha_0)^{-1}n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha_0) + o_{\text{pr}}(1/\sqrt{n}).$$

Let H_{xy} be the matrix obtained by first differentiating $\text{E}\log[1 + \lambda^T M(\mu)]$ with respect to y and afterwards with respect to x evaluated at (λ_0, μ_0) . Then $J\Gamma^{\text{EL}}(\alpha_0)$ can be written

as the following block matrix

$$J\Gamma^{\text{EL}}(\alpha_0) = \begin{pmatrix} H_{\lambda\lambda} & H_{\mu\lambda} \\ H_{\lambda\mu} & H_{\mu\mu} \end{pmatrix}.$$

Furthermore, $\lambda'(\mu_0) = -H_{\lambda\lambda}^{-1}H_{\mu\lambda}$ by the implicit function theorem. Utilizing this and the formula for inverses of block matrices, we find $(0, I)\Gamma^{\text{EL}}(\alpha_0) = (A^{-1}\lambda'(\mu_0)^T, A^{-1}) = A^{-1}(\lambda'(\mu_0)^T, I)$, where $A = H_{\mu\mu} - H_{\lambda\mu}H_{\lambda\lambda}^{-1}H_{\mu\lambda}$. Direct computation shows that $A = J_{\text{EL}}$. Putting all of this together shows $(0, I)\Gamma^{\text{EL}}(\alpha_0)^{-1}\psi(y, \alpha_0) = J_{\text{EL}}^{-1}(\lambda'(\mu_0)^T, I)\psi(y, \lambda_0, \mu_0) = J_{\text{EL}}^{-1}\psi_\mu(y, \mu_0)$, proving (6) in the main part of the article.

2 Proofs for Section 3

As with Theorem 2.1 we will first show a slightly different result.

Theorem 2.1. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last s arguments. Assume that m is used in construction of the empirical likelihood function EL_n . Let f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$ be some parametric family with score function u and $\mu: \mathbb{R}^p \rightarrow \mathbb{R}^s$ such that $\text{Em}[Y, \mu(\theta)] = 0$ when $Y \sim f_\theta$. Fix $a \in [0, 1)$ and assume that $\hat{\theta}$ is the unique maximizer of h_n . Furthermore, let $\hat{\lambda}(\hat{\theta})$ be the unique solution to*

$$0 = n^{-1} \sum_{i=1}^n \frac{m[Y_i, \mu(\hat{\theta})]}{1 + \lambda^T m[Y_i, \mu(\hat{\theta})]} = 0$$

Lastly, let $\beta = (\theta, \lambda)$, $\hat{\beta} = (\hat{\theta}, \hat{\lambda})$, and

$$\psi_{\text{HL}}(y, \beta) = (1 - a) \begin{pmatrix} 0 \\ u(y, \theta) \end{pmatrix} - a \begin{pmatrix} I & 0 \\ 0 & \mu'(\theta)^T \end{pmatrix} \psi_{\text{EL}}[y, \lambda, \mu(\theta)].$$

Then $\hat{\beta}$ is consistent for $\beta_0 = (\lambda_0, \mu_0)$, the unique root of $\Gamma^{\text{HL}}(\lambda, \mu) = \mathbb{E}\psi(Y, \beta)$, where $Y \sim F$. Furthermore,

$$\hat{\beta} = \beta_0 - n^{-1} J \Gamma^{\text{EL}}(\beta_0)^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (2)$$

provided condition (1-3) of Theorem 1.2 hold true with $\alpha_0 = (\lambda_0^T, \mu(\theta_0)^T)^T$ and the following additional assumptions

(4) The function $\mu(\theta)$ is twice continuously differentiable in a neighborhood of θ_0 .

(5) There exists a neighborhood of θ_0 on which all components of u and $\partial u / \partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

In particular, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, J \Gamma^{\text{HL}}(\beta_0)^{-1} K_{\text{HL}} J \Gamma^{\text{HL}}(\beta_0)^{-1})$, with $K_{\text{HL}} = \text{Var}\psi_{\text{HL}}(Y, \beta_0)$, $Y \sim F$.

Proof. Notice that

$$\begin{aligned} \frac{\partial h_n(\theta)}{\partial \theta} &= \sum_{i=1}^n \left\{ (1-a)u(Y_i, \theta) - a\mu'(\theta)^T \frac{\partial m[Y_i, \mu(\theta)]^T \lambda_n(\theta) + \lambda_n'(\theta)^T m[Y_i, \mu(\theta)]}{1 + \lambda_n(\theta)^T m[Y_i, \mu(\theta)]} \right\} \\ &= \sum_{i=1}^n \left\{ (1-a)u(Y_i, \theta) - a\mu'(\theta)^T \frac{\partial m[Y_i, \mu(\theta)]^T \lambda_n(\theta)}{1 + \lambda_n(\theta)^T m[Y_i, \mu(\theta)]} \right\}, \end{aligned}$$

where the last equality follows from the definition of $\lambda_n(\theta)$ and ∂m is short hand for $\partial m / \partial \mu$.

Hence, $\hat{\beta}$ is the solution to $0 = n^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta)$ where ψ_{HL} is as defined in the theorem. Because of this $\hat{\beta}$ is a Z-estimator and, just as for the MEL estimator, standard results can be used to study its limiting behavior.

Let $m_2: \mathbb{R}^{d+p} \rightarrow \mathbb{R}^q$ be defined as $m_2(y, \theta) = m[y, \mu(\theta)]$. Since μ is twice continuously differentiable as a function of θ , direct computations shows that conditions (1-3) hold for m_2 . By the proof of Theorem 1.2 this ensures that there exists a neighborhood on which

the norm of

$$\begin{pmatrix} I & 0 \\ 0 & \mu'(\theta)^T \end{pmatrix} \psi_{\text{EL}}[y, \lambda, \mu(\theta)] = \begin{pmatrix} -m_2(y, \theta) \\ -\partial m_2(y, \theta)^T \lambda \end{pmatrix} [1 + \lambda^T m_2(y, \theta)]^{-1} \quad (3)$$

is bounded some square F -integrable function depending on neither λ nor μ . Combined with assumption (5), this proves that the norm of $\psi(y, \beta)$ is bounded by some square integrable function not depending on β . In particular, (C2) of Lemma 1.1 holds true. As (C1) holds by assumption, consistency of $\hat{\beta}$ towards β_0 follows.

To show (2), we will again use Theorem 5.21 in Van der Vaart (2000). The only condition which remains to show is the Lipschitz condition. By the proof of Theorem 1.2, the norm of all partial derivatives of (3) are bounded by some square integrable function. Combining this with assumption (5) of this theorem and the mean value theorem, ensures that there exists a square integrable function r such that $\|\psi_{\text{HL}}(y, \beta_1) - \psi_{\text{HL}}(y, \beta_2)\| \leq r(y)\|\beta_1 - \beta_2\|$. This proves the Lipschitz condition of Theorem 5.21 in Van der Vaart (2000) and concludes the proof. $\square$

The proof of Theorem 3.1 now follows directly from Theorem 2.1 by performing computations similar to those in Section 1.1.

References

- Imbens, G. W., Spady, R. H. & Johnson, P. (1998), ‘Information theoretic approaches to inference in moment condition models’.
- Joe, H. (2005), ‘Asymptotic efficiency of the two-stage estimation method for copula-based models’, *Journal of multivariate Analysis* **94**(2), 401–419.

- Schennach, S. M. (2007), ‘Point estimation with exponentially tilted empirical likelihood’, *The Annals of Statistics* **35**, 634–672.
- Van der Vaart, A. W. (2000), *Asymptotic statistics*, Cambridge University Press, Cambridge, England.