APPENDIX

A Main results

In this section of the appendix we provide the proofs for the main results of Section 2. As in the rest of the paper, we always implicitly assume that all functions of y are measurable, and that all expectations and integrals over y are well-defined.

A.1 Proof of Theorem 1

A.1.1 Notation and assumptions

In all our applications Π is either a vector space or an affine space. Let $\overline{\mathcal{T}}$ and $\mathcal{T}$ be the tangent and cotangent spaces of Π at $\pi(\gamma_*)$. Thus, for $\pi_1, \pi_2 \in \Pi$ we have $(\pi_1 - \pi_2) \in \overline{\mathcal{T}}$, and $\mathcal{T}$ is the set of linear maps $u : \overline{\mathcal{T}} \to \mathbb{R}$. For a scalar function $q : \Pi \mapsto \mathbb{R}$, we have $\nabla_\pi q_{\pi(\gamma_*)} \in \mathcal{T}$; that is, the typical element of $\mathcal{T}$ is a gradient. Conversely, for a map to Π, such as $\gamma \mapsto \pi(\gamma)$, we have $\frac{\partial \pi(\gamma_*)}{\partial \gamma_k} \in \overline{\mathcal{T}}$.

For $v \in \overline{\mathcal{T}}$ and $u \in \mathcal{T}$ we use the bracket notation $\langle v, u \rangle \in \mathbb{R}$ to denote the bilinear mapping. Here we are not assuming a Hilbert space structure, and we only use the bracket notation to combine vectors v and covectors u into a scalar.

Our squared distance measure $d(\pi_0, \pi(\gamma_*))$ on Π induces a norm on the tangent space $\overline{\mathcal{T}}$, namely for $v \in \overline{\mathcal{T}}$,

$$\|v\|^2_{\text{ind},\gamma_*} = \lim_{\epsilon \to 0} \frac{d\big(\pi(\gamma_*) + \epsilon^{1/2} v, \pi(\gamma_*)\big)}{\epsilon}.$$

Throughout we assume that $\dim \beta$ and $\dim \gamma$ are finite. For any finite-dimensional vectors we use the standard Euclidean norm $\|\cdot\|$, and for any finite-dimensional matrices we use the spectral matrix norm, which we also denote by $\|\cdot\|$. Let $\mathcal{Y}$ denote the support of Y.

Assumption A1. *We assume that $Y_i \sim$ i.i.d.f_{β_0,π_0}. In addition, we impose the following regularity conditions:*

(i) *We consider $n \to \infty$ and $\epsilon \to 0$ such that $\epsilon n \to c$, for some constant $c \in (0, \infty)$.*

(ii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\| \nabla_\pi \delta_{\beta_0,\pi_0} \right\|_{\gamma_*} = O(1)$, *and*
$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} \right\rangle \right| = o(\epsilon^{1/2})$.

(iii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right]^2 dy \right\}^{1/2} = O(\epsilon^{1/2})$,
$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|^2_{\gamma_*} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right]^2 dy = o(1)$,
$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right\rangle \right]^2 dy = o(\epsilon)$.

(iv) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \epsilon^{-1/2} \|\pi_0 - \pi(\gamma_*)\|_{\text{ind},\gamma_*} = 1 + o(1)$. *Furthermore, for* $u_\epsilon \in \mathcal{T}$ *with* $\|u_\epsilon\|_{\gamma_*} = O(1)$ *we have*

$$\left| \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \epsilon^{-1/2} \langle \pi_0 - \pi(\gamma_*), u_\epsilon \rangle - \|u_\epsilon\|_{\gamma_*} \right| = o(1).$$

(v) *For some* $\nu > 0$ *we have* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left\| \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|^{2+\nu} = O(1)$, *and* $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|_{\gamma_*}^{2+\nu} = O(1)$.
Furthermore we assume that $\left\| \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)} \right\| = O(1)$, *and* $\left\| H_{\beta\gamma}^{-1} \right\| = O(1)$.

Part (i) of Assumption A1 describes our asymptotic framework, where the assumption $\epsilon n \to c$ is required to ensure that the squared worst-case bias (of order ϵ) and the variance (of order $1/n$) of the estimators for δ_{β_0,π_0} are asymptotically of the same order, so that the MSE provides a meaningful balance between bias and variance asymptotically. Part (ii) requires δ_{β_0,π_0} to be sufficiently smooth in π_0, so that a first-order Taylor expansion provides a good local approximation to δ_{β_0,π_0}.

Part (iii) of Assumption A1 is a smoothness assumption on $f_{\beta_0,\pi_0}(y)$ in π_0. Those conditions may not look intuitive, in particular when π_0 is infinite-dimensional, so we want to discuss that assumption in some more detail here for the case of the semi-parametric mixture models introduced in Section 3.4, where$f_{\beta_0,\pi_0}(y) = \int_{\mathcal{A}} g_{\beta_0}(y|a)\, \pi_0(a)\, da$. In that case we have

$$\begin{aligned} \int_{\mathcal{Y}} \left[f_{\beta_0,\pi_0}^{1/2}(y) - f_{\beta_0,\pi(\gamma_*)}^{1/2}(y) \right]^2 dy &= 2\, H^2(f_{\beta_0,\pi_0}, f_{\beta_0,\pi(\gamma_*)}) \\ &\leq 2\, D_{\text{KL}}(f_{\beta_0,\pi_0} || f_{\beta_0,\pi(\gamma_*)}) \leq 2\, D_{\text{KL}}(\pi_0 || \pi(\gamma_*)), \end{aligned}$$

where the first inequality is the general relation $H^2(f_{\beta_0,\pi_0}, f_{\beta_0,\pi(\gamma_*)}) \leq D_{\text{KL}}(f_{\beta_0,\pi_0} || f_{\beta_0,\pi(\gamma_*)})$ between the squared Hellinger distance H^2 and the Kullback-Leibler divergence D_{KL}, and the second inequality is sometimes called the "chain rule" for the Kullback-Leibler divergence, which can be derived by an application of Jensen's inequality. Since we defined our distance measure $d(\pi_0, \pi(\gamma_*))$ in the semi-parametric mixture case to be twice the Kullback-Leibler divergence $2D_{\text{KL}}(\pi_0 || \pi(\gamma_*)) = 2\, \mathbb{E}_{\pi_0} \log[\pi_0(A)/\pi(A \mid \gamma_*)]$ we find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ \int_{\mathcal{Y}} \left[f_{\beta_0,\pi_0}^{1/2}(y) - f_{\beta_0,\pi(\gamma_*)}^{1/2}(y) \right]^2 dy \right\}^{1/2} \leq \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \{ d(\pi_0, \pi(\gamma_*)) \}^{1/2} = \epsilon^{1/2}.$$

Thus, the first condition in Assumption A1(iii) is satisfied for those semi-parametric mixture models.

The second condition in Assumption A1(iii) can be justified by imposing that

$$\sup_{y \in \mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|_{\gamma_*}^2 = O(1),$$

which for the semi-parametric mixture model can equivalently be written as

$$\sup_{y \in \mathcal{Y}} \frac{\text{Var}_{\pi(\gamma_*)} \left[g_{\beta_0}(y \mid A) \right]}{\left[\mathbb{E}_{\pi(\gamma_*)}\, g_{\beta_0}(y \mid A) \right]^2} = O(1). \tag{A1}$$

For any standard discrete choice model (as those discussed in Section 6) we have that $\sup_{y\in\mathcal{Y}} \mathrm{Var}_{\pi(\gamma_*)}\left[g_{\beta_0}(y\,|\,A)\right] < \infty$, and $\inf_{y\in\mathcal{Y}} \mathbb{E}_{\pi(\gamma_*)}\, g_{\beta_0}(y\,|\,A) > 0$, implying that equation (A1) is satisfied. We then have

$$\begin{aligned}
&\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left\|\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y)\right\|^2_{\gamma_*} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y)\right]^2 dy \\
&\quad\le \underbrace{\left[\sup_{y\in\mathcal{Y}} \left\|\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y)\right\|^2_{\gamma_*}\right]}_{=O(1)} \underbrace{\left\{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y)\right]^2 dy\right\}}_{\le\epsilon=o(1)} = o(1).
\end{aligned}$$

Thus, one way to justify the second condition in Assumption A1(iii) is to argue that equation (A1) holds, which is the case for our illustrations in Section 6. The last condition in Assumption A1(iii) could be broken down analogously for semi-parametric mixture models, but it is actually a standard condition of differentiability in quadratic mean that is also regularly imposed when π is infinite-dimensional (see, e.g., equation (5.38) in Van der Vaart, 2007).

Part (iv) of Assumption A1 requires that our distance measure $d(\pi_0, \pi(\gamma_*))$ converges to the associated norm for small values ϵ in a smooth fashion. Finally, part (v) requires invertibility of $H_{\beta\gamma}$ (but invertibility of H_π or $\widetilde{H}_\pi$ are *not* required), uniform boundedness of various derivatives, and of the $(2+\nu)$-th moment of $\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y)$ — which again can be justified by equation (A1), because we then have $\sup_{y\in\mathcal{Y}} \left\|\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y)\right\|^2_{\gamma_*} = O(1)$.

For many of the proofs we only need the regularity conditions in Assumption A1. However, in order to describe the properties of our minimum-MSE estimator $\widehat{\delta}^{\,\mathrm{MMSE}}_{\epsilon} = \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^n h^{\mathrm{MMSE}}_\epsilon(Y_i, \widehat{\beta}, \widehat{\gamma})$ we also need to account for the fact that $\widehat{\beta}$ and $\widehat{\gamma}$ themselves are estimated. It turns our that the leading-order asymptotic properties of $\widehat{\delta}^{\,\mathrm{MMSE}}_{\epsilon}$ are independent of whether β_0 and γ_* are known or estimated in the construction of $\widehat{\delta}^{\,\mathrm{MMSE}}_{\epsilon}$ (see, e.g., Lemma A3 below), but formally showing this requires some additional assumptions, which we present next.

Assumption A2. *For some $\chi > 2$ we have*

(i) $$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} \left\| \begin{pmatrix}\widehat{\beta}\\ \widehat{\gamma}\end{pmatrix} - \begin{pmatrix}\beta_0\\ \gamma_*\end{pmatrix} \right\|^\chi\right)^{\frac{1}{\chi}} = O\left(\frac{1}{\sqrt{n}}\right).$$

(ii) $$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left\|\nabla_\eta\, h^{\mathrm{MMSE}}_\epsilon(Y, \beta_0, \gamma_*)\right\| = O(1),\ \textit{where}\ \eta = (\beta', \gamma')'.$$

(iii) $$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \sup_{\beta\in\mathcal{B},\gamma\in\mathcal{G}} \left\| \frac{1}{n}\sum_{i=1}^n \nabla^2_{\eta\eta'} h^{\mathrm{MMSE}}_\epsilon(Y_i, \beta, \gamma)\right\| = O(1),\ \textit{where}\ \eta = (\beta', \gamma')'.$$

(iv) $$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[h^{\mathrm{MMSE}}_\epsilon(Y, \beta_0, \gamma_*)\right]^{2+\nu} = O(1),\ \textit{for some}\ \nu > 0.$$

Part (i) of Assumption A2 requires $\widehat{\beta}$ and $\widehat{\gamma}$ to converge at $\sqrt{n}$ rate. As discussed in the main text, we assume that preliminary estimators have finite χ-moments where $\chi > 2$. Part

(ii) of Assumption A2 requires a uniformly bounded second moment for $\nabla_\eta h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$. Since equation (20) in the main text gives an explicit expression for $h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$, we could replace Assumption A2(ii) by appropriate assumptions on the model primitives $f_{\beta_0,\pi_0}(y)$ and δ_{β_0,π_0}, but for the sake of brevity we state the assumption in terms of $h_\epsilon^{\text{MMSE}}(y, \beta_0, \gamma_*)$. The same is true for part (iii) of Assumption A2. Notice that this last part of the assumption involves a supremum over β and γ inside of an expectation – in order to verify it, one either requires a uniform Lipschitz bound on the dependence of $h_\epsilon^{\text{MMSE}}(Y_i, \beta, \gamma)$ on β and γ, or some empirical process method to control the entropy of that function (e.g., a bracketing argument). But since β and γ are finite-dimensional parameters these are all standard arguments.

We verified Assumption A2(iv) in the locally quadratic case of Section 3. Formally, we have the following lemma.

Lemma A1. *Let Assumption A1 hold, and assume that $h_\epsilon^{\text{MMSE}}(\cdot, \beta_0, \gamma_*)$ is given by Lemma 1 in the main text. Then, Assumption A2(iv) holds with the constant ν specified in Assumption A1.*

A.1.2 Proof of Theorem 1

For a function $h_\epsilon = h_\epsilon(y, \beta_0, \gamma_*)$ we define

$$\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) := \delta_{\beta_0,\pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^{n} h_\epsilon(Y_i, \beta_0, \gamma_*).$$

It is useful to establish some preliminary lemmas before showing the main result. The proofs for those lemmas are provided in Section S1.1.

Lemma A2. *Let Assumption A1 hold, and let $h_\epsilon(\cdot, \beta_0, \gamma_*)$ be a sequence of influence functions that satisfy the unbiasedness constraint (2) as well as $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$, for some $\kappa > 2$. Then,*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0}\right]^2 = b_\epsilon(h_\epsilon, \beta_0, \gamma_*)^2 + \frac{\operatorname{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y, \beta_0, \gamma_*))}{n} + o(\epsilon).$$

Lemma A2 provides a formal justification for the worst-case MSE approximation introduced in equation (9) of the main text.

Recall that $\widehat{\delta}_\epsilon^{\text{MMSE}} = \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \widehat{\beta}, \widehat{\gamma})$. This differs from $\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*)$, because β_0 and γ_* have to be estimated. The following lemma shows that the fact that β_0 and γ_* are estimated in the construction of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ can be neglected to first order. Notice that this result requires the additional regularity conditions in Assumption A2, which are not required anywhere else in the proof of Theorem 1.

Lemma A3. *Let Assumptions A1 and A2 hold. Then,*

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*)\right| = O\left(\frac{1}{n}\right).$$

Thus, Lemma A3 guarantees that $\widehat{\delta}_\epsilon^{\text{MMSE}} = \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) + O_{P_0}(1/n)$. This may be surprising given that the differences $\widehat{\beta} - \beta_0$ and $\widehat{\gamma} - \gamma_*$ are themselves of order $1/\sqrt{n}$. However, recall that by construction h_ϵ^{MMSE} satisfies the local robustness condition (3), which is imposed through our constraints (2) and (4). Local robustness ensures that $\widehat{\beta} - \beta_0$ and $\widehat{\gamma} - \gamma_*$ have no leading-order effect on $\widehat{\delta}_\epsilon^{\text{MMSE}} - \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*)$.

For the next lemma, recall the decomposition of $\widehat{\delta}_\epsilon$ in Theorem 1 in the main text:

$$\begin{aligned}\widehat{\delta}_\epsilon &= \delta_{\beta_0,\pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + n^{-1/2} R_n \\ &= \widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) + n^{-1/2} R_n. \end{aligned} \tag{A2}$$

Here, $\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*)$ is the well-behaved leading-order contribution to $\widehat{\delta}_\epsilon$, whereas R_n is an asymptotically vanishing remainder term that may, however, have heavy tails (it only satisfies a trimmed second moment condition). The following lemma shows that the worst-case trimmed MSE of $\widehat{\delta}_\epsilon$ is bounded from below by the MSE of the leading-order term $\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*)$.

Lemma A4. *Let Assumption A1 hold, and let $h_\epsilon(\cdot, \beta_0, \gamma_*)$ be a sequence of influence functions that satisfy the unbiasedness constraint (2) as well as $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$, for some $\kappa > 2$. Assume that (A2) holds, and let $m_n > 0$ be a sequence such that $m_n\, n^{1/2}\, [\log(n)]^{-1} \to \infty$. Furthermore, assume that*

(i) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathrm{P}_{\beta_0,\pi_0}(|R_n| > \log(n)) = o(1)$,

(ii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\, \mathbb{1}(|R_n| \le 2\log(n))\right] = o(1)$.

Then we have

$$\begin{aligned}&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0}\right)^2\right] \\ &\quad \le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] + o(\epsilon). \end{aligned} \tag{A3}$$

We now have all the preliminary results required to show the main theorem.

Proof of Theorem 1. Define

$$r_\epsilon := \widehat{\delta}_\epsilon^{\text{MMSE}} - \widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*).$$

We then have

$$\begin{aligned}&\mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] \\ &= \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}}, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0} + r_\epsilon\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\beta_0,\pi_0}\Bigg[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2 \underbrace{\mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)}_{\le 1} \Bigg] \\
&\quad + 2\,\mathbb{E}_{\beta_0,\pi_0}\left[r_\epsilon \left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0} + r_\epsilon\right) \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] \\
&\quad \underbrace{-\mathbb{E}_{\beta_0,\pi_0}\left[r_\epsilon^2\, \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right]}_{\le 0} \\
&\le \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2 \right] \\
&\quad + 2\,\mathbb{E}_{\beta_0,\pi_0}\Bigg[\underbrace{r_\epsilon \left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right) \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)}_{\le |r_\epsilon|\, m_n} \Bigg] \\
&\le \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2 \right] + 2\, m_n\, \mathbb{E}_{\beta_0,\pi_0} \left|r_\epsilon\right| .
\end{aligned}
$$

According to Lemma A3 we have $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |r_\epsilon| = O(1/n) = O(\epsilon)$, and the assumptions of the theorem guarantee that $m_n = o(1)$. We thus obtain

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] \\
&\qquad \le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2 \right] + o(\epsilon). \qquad \text{(A4)}
\end{aligned}
$$

By definition h_ϵ^{MMSE} also satisfies the unbiasedness constraint (2). Together with Assumption A2(iv) this implies that h_ϵ^{MMSE} satisfies the conditions on h_ϵ in Lemma A2 with $\kappa = 2+\nu$. Thus, we can apply Lemma A2 with $h_\epsilon = h_\epsilon^{\text{MMSE}}$ to find that

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right)^2 \right] \\
&\qquad = b_\epsilon(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*))}{n} + o(\epsilon). \qquad \text{(A5)}
\end{aligned}
$$

The function $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ is defined by the minimization problem (10) in the main text. In other words, $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ minimizes the objective function $b_\epsilon(h,\beta_0,\gamma_*)^2 + n^{-1}\text{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,\beta_0,\gamma_*))$, subject to the constraints (2) and (4). Theorem 1 assumes that $h_\epsilon = h_\epsilon(\cdot,\beta_0,\gamma_*)$ satisfies those constraints, and the definition of $h_\epsilon^{\text{MMSE}}(\cdot,\beta_0,\gamma_*)$ therefore implies that

$$
\begin{aligned}
&b_\epsilon(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon^{\text{MMSE}}(Y,\beta_0,\gamma_*))}{n} \\
&\qquad \le b_\epsilon(h_\epsilon,\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*))}{n}. \qquad \text{(A6)}
\end{aligned}
$$

Theorem 1 also imposes all the assumptions on h_ϵ in Lemma A2 . By applying that lemma we thus have

$$
b_\epsilon(h_\epsilon,\beta_0,\gamma_*)^2 + \frac{\text{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*))}{n}
$$

$$= \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}(h_\epsilon, \beta_0, \gamma_*) - \delta_{\beta_0,\pi_0}\right)^2\right] + o(\epsilon). \tag{A7}$$

Finally, Theorem 1 also guarantees all the assumptions of Lemma A4, implying that the inequality (A3) holds. Now, combining (A4), (A5), (A6), (A7) and (A3) gives

$$\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon^{\text{MMSE}} - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] \\
&\qquad \le \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\left(\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0}\right)^2 \mathbb{1}\left(\left|\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0}\right| \le m_n\right)\right] + o(\epsilon),
\end{aligned} \tag{A8}$$

which is what we wanted to show. ∎

A.2 Proof of Theorem 2

Assumption A3.

(i) $\widehat{\delta} - \delta_{\beta_0,\pi(\gamma_*)} - \frac{1}{n}\sum_{i=1}^n h(Y_i, \beta_0, \gamma_*) = o_{P_{\beta_0,\pi_0}}(n^{-\frac{1}{2}})$, *uniformly in* $\pi_0 \in \Gamma_\epsilon(\gamma_*)$.

(ii) Let $\sigma_h^2(\beta_0, \pi_0, \gamma_*) = \text{Var}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)$. *We assume that there exists a constant* c, *independent of* ϵ, *such that* $\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \sigma_h(\beta_0, \pi_0, \gamma_*) \ge c > 0$. *Furthermore, for all sequences* $a_n = c_{1-\alpha/2} + o(1)$ *we have*

$$\inf_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \text{Pr}_{\beta_0,\pi_0}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)}\right| \le a_n\right] \ge 1 - \alpha + o(1).$$

(iii) $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\|\widehat{\beta} - \beta_0\|^2 = o(1)$, $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\|\widehat{\gamma} - \gamma_*\|^2 = o(1)$, $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}[\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)]^2 = o(1)$.

(iv) $\|\nabla_{\beta\gamma} b_\epsilon(h, \beta, \gamma)\| = O(\epsilon^{\frac{1}{2}})$, *uniformly in some neighborhood around* β_0, γ_*.

Part (i) is weaker than the local regularity of the estimator $\widehat{\delta}$ that we assumed when analyzing the minimum-MSE estimator; see equation (14). In turn, related to but differently from the conditions we used for Theorem 1, part (ii) requires a form of local asymptotic normality of the estimator.

Proof of Theorem 2. Let $\widehat{\delta}$ be an estimator and $h(y, \beta_0, \gamma_*)$ be the corresponding influence function such that part (i) in Assumption A3 holds. Define $\widehat{R}_{\beta_0,\gamma_*} := \widehat{\delta} - \delta_{\beta_0,\pi(\gamma_*)} - \frac{1}{n}\sum_{i=1}^n h(Y_i, \beta_0, \gamma_*)$. We then have

$$\begin{aligned}
\widehat{\delta} - \delta_{\beta_0,\pi_0} &= \frac{1}{n}\sum_{i=1}^n h(Y_i, \beta_0, \gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \widehat{R}_{\beta_0,\gamma_*} \\
&= \frac{1}{n}\sum_{i=1}^n \left[h(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)\right] - \left[\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*)\right] + \widehat{R}_{\beta_0,\gamma_*},
\end{aligned}$$

and therefore

$$\underbrace{\frac{|\widehat{\delta}-\delta_{\beta_0,\pi_0}|-b_\epsilon(h,\widehat{\beta},\widehat{\gamma})-\widehat{\sigma}_h\, c_{1-\alpha/2}/\sqrt{n}}{\sigma_h(\beta_0,\pi_0,\gamma_*)/\sqrt{n}}}_{=\text{lhs}} \leq \underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{h(Y_i,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi_0}h(Y,\beta_0,\gamma_*)}{\sigma_h(\beta_0,\pi_0,\gamma_*)}\right|-c_{1-\alpha/2}+\widehat{r}_{\beta_0,\pi_0,\gamma_*}}_{=\text{rhs}}, \tag{A9}$$

where

$$\begin{aligned}
&\widehat{r}_{\beta_0,\pi_0,\gamma_*}\\
&:= c_{1-\alpha/2}+\frac{\left|\delta_{\beta_0,\pi_0}-\delta_{\beta_0,\pi(\gamma_*)}-\mathbb{E}_{\beta_0,\pi_0}h(Y,\beta_0,\gamma_*)\right|+\left|\widehat{R}_{\beta_0,\gamma_*}\right|-b_\epsilon(h,\widehat{\beta},\widehat{\gamma})-\widehat{\sigma}_h\, c_{1-\alpha/2}/\sqrt{n}}{\sigma_h(\beta_0,\pi_0,\gamma_*)/\sqrt{n}}\\
&=\frac{\sqrt{n}}{\sigma_h(\beta_0,\pi_0,\gamma_*)}\Bigg\{|\delta_{\beta_0,\pi_0}-\delta_{\beta_0,\pi(\gamma_*)}-\mathbb{E}_{\beta_0,\pi_0}h(Y,\beta_0,\gamma_*)|+|\widehat{R}_{\beta_0,\gamma_*}|\\
&\qquad -b_\epsilon(h,\widehat{\beta},\widehat{\gamma})-\frac{\widehat{\sigma}_h-\sigma_h(\beta_0,\pi_0,\gamma_*)}{\sqrt{n}}c_{1-\alpha/2}\Bigg\}.
\end{aligned}$$

From (A9), we conclude that the event rhs ≤ 0 implies the event lhs ≤ 0, and therefore $\Pr_{\beta_0,\pi_0}(\text{lhs}\leq 0)\geq \Pr_{\beta_0,\pi_0}(\text{rhs}\leq 0)$, which we can also write as

$$\begin{aligned}
&\Pr_{\beta_0,\pi_0}\left[|\widehat{\delta}-\delta_{\beta_0,\pi_0}|\leq b_\epsilon(h,\widehat{\beta},\widehat{\gamma})+\frac{\widehat{\sigma}_h}{\sqrt{n}}c_{1-\alpha/2}\right]\\
&\qquad\geq \Pr_{\beta_0,\pi_0}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{h(Y_i,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi_0}h(Y,\beta_0,\gamma_*)}{\sigma_h(\beta_0,\pi_0,\gamma_*)}\right|\leq c_{1-\alpha/2}-\widehat{r}_{\beta_0,\pi_0,\gamma_*}\right].
\end{aligned} \tag{A10}$$

By part (iv) in Assumption A3 there exists a constant $C>0$ such that $\|\nabla_{\beta\gamma}b_\epsilon(h,\beta,\gamma)\|\leq C\,\epsilon^{\frac{1}{2}}$, uniformly in a neighborhood of (β_0,γ_*), and therefore

$$\left|b_\epsilon(h,\widehat{\beta},\widehat{\gamma})-b_\epsilon(h,\beta_0,\gamma_*)\right|\leq C\,\epsilon^{\frac{1}{2}}\left\|\begin{pmatrix}\widehat{\beta}-\beta_0\\ \widehat{\gamma}-\gamma_*\end{pmatrix}\right\|.$$

Using this we find that

$$\begin{aligned}
\left|\widehat{r}_{\beta_0,\pi_0,\gamma_*}\right|\leq \frac{\sqrt{n}}{\sigma_h(\beta_0,\pi_0,\gamma_*)}&\Bigg\{\Big|\left|\delta_{\beta_0,\pi_0}-\delta_{\beta_0,\pi(\gamma_*)}-\mathbb{E}_{\beta_0,\pi_0}h(Y,\beta_0,\gamma_*)\right|-b_\epsilon(h,\beta_0,\gamma_*)\Big|\\
&+\frac{|\widehat{\sigma}_h-\sigma_h(\beta_0,\pi_0,\gamma_*)|}{\sqrt{n}}c_{1-\alpha/2}+C\,\epsilon^{\frac{1}{2}}\left\|\begin{pmatrix}\widehat{\beta}-\beta_0\\ \widehat{\gamma}-\gamma_*\end{pmatrix}\right\|+|\widehat{R}_{\beta_0,\gamma_*}|\Bigg\}.
\end{aligned}$$

Parts (i) and (ii) of Assumption A3 imply that, uniformly in $\pi_0\in\Gamma_\epsilon(\gamma_*)$, we have

$$\frac{\sqrt{n}}{\sigma_h(\beta_0,\pi_0,\gamma_*)}\widehat{R}_{\beta_0,\gamma_*}=o_{P_{\beta_0,\pi_0}}(1),$$

and analogously we find from the conditions in Assumption A3 that

$$\frac{\widehat{\sigma}_h - \sigma_h(\beta_0, \pi_0, \gamma_*)}{\sigma_h(\beta_0, \pi_0, \gamma_*)} = o_{P_{\beta_0,\pi_0}}(1), \qquad \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \epsilon^{\frac{1}{2}} \left\| \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\gamma} - \gamma_* \end{pmatrix} \right\| = o_{P_{\beta_0,\pi_0}}(1),$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Finally, since we also impose Assumption A1 and $\sup_{\pi_0 \in \Gamma_\epsilon} \mathbb{E}_{\beta_0,\pi_0} h^2(Y, \beta_0, \gamma_*) = O(1)$ we obtain, analogously to the proof of Lemma S1(iii) in Section S1, that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \frac{\sqrt{n}}{\sigma_h(\beta_0, \pi_0, \gamma_*)} \left| \left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} - \mathbb{E}_{\beta_0,\pi_0} h(Y, \beta_0, \gamma_*) \right| - b_\epsilon(h, \beta_0, \gamma_*) \right| = o(1).$$

We thus conclude that $\widehat{r}_{\beta_0,\pi_0,\gamma_*} = o_{P_{\beta_0,\pi_0}}(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Together with (A10) and part (ii) in Assumption A3 this implies (17), hence Theorem 2. ∎

SUPPLEMENTARY APPENDIX

"Minimizing Sensitivity to Model Misspecification"

Stéphane Bonhomme and Martin Weidner

In Sections S1 and S2, we provide details about the proofs in the paper. In Section S3, we describe our computational approach. In Section S4, we outline how to extend our approach to models defined by moment restrictions. Lastly, we report additional simulation and estimation results in Section S5.

S1 Complements to main results of Section 2

S1.1 Proof of intermediate lemmas for Theorem 1

The proofs of the Lemmas A2, A3 and A4 are provided in this subsection. Before those proofs it is useful to first establish one additional lemma.

Lemma S1. *Let Assumption A1 hold. Let $q_\epsilon(y)$ and $h_\epsilon(y,\beta_0,\gamma_*)$ be sequences of functions with $\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\left|q(Y)\right|^\zeta = O(1)$, for some $\zeta>1$, and $\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\left|h_\epsilon(Y,\beta_0,\gamma_*)\right|^2 = O(1)$. Then we have*

$$\text{(i)}\quad \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left|\delta_{\beta_0,\pi(\gamma_*)}-\delta_{\beta_0,\pi_0}\right| = O(\epsilon^{1/2}),$$

$$\text{(ii)}\quad \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left|\mathbb{E}_{\beta_0,\pi_0}q_\epsilon(Y)-\mathbb{E}_{\beta_0,\pi(\gamma_*)}q_\epsilon(Y)\right| = O(\epsilon^{1/2}),$$

$$\text{(iii)}\quad \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\Big|\mathbb{E}_{\beta_0,\pi_0}h_\epsilon(Y,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi(\gamma_*)}h_\epsilon(Y,\beta_0,\gamma_*) - \left\langle \pi_0-\pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)}h_\epsilon(Y,\beta_0,\gamma_*)\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y)\right\rangle\Big| = o(\epsilon^{1/2}).$$

Proof of Lemma S1. # Part (i): By a mean-value expansion around $\pi(\gamma_*)$ we find

$$\left|\delta_{\beta_0,\pi_0}-\delta_{\beta_0,\pi(\gamma_*)}\right| = \left|\left\langle \pi_0-\pi(\gamma_*),\nabla_\pi\delta_{\beta_0,\widetilde{\pi}}\right\rangle\right| \leq \left\|\pi_0-\pi(\gamma_*)\right\|_{\mathrm{ind},\gamma_*}\left\|\nabla_\pi\delta_{\beta_0,\widetilde{\pi}}\right\|_{\gamma_*},$$

where $\widetilde{\pi}$ is between $\pi(\gamma_*)$ and π_0. Therefore

$$\begin{aligned}\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left|\delta_{\beta_0,\pi_0}-\delta_{\beta_0,\pi(\gamma_*)}\right| &\leq \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left\|\pi_0-\pi(\gamma_*)\right\|_{\mathrm{ind},\gamma_*}\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left\|\nabla_\pi\delta_{\beta_0,\pi_0}\right\|_{\gamma_*}\\ &= O(\epsilon^{1/2})\,O(1) = O(\epsilon^{1/2}).\end{aligned}$$

Part (ii): Without loss of generality we assume that $\zeta\leq 2$. Let $\xi := \zeta/(\zeta-1)\geq 2$. We then have

$$\int_{\mathcal{Y}}\left|f^{1/\xi}_{\beta_0,\pi_0}(y)-f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)\right|^\xi dy \leq \int_{\mathcal{Y}}\left[f^{1/2}_{\beta_0,\pi_0}(y)-f^{1/2}_{\beta_0,\pi(\gamma_*)}(y)\right]^2 dy,$$

where we used that $|a-b| \leq |a^c - b^c|^{1/c}$, for any $a, b \geq 0$ and $c \geq 1$, and plugged in $a = f^{1/\xi}_{\beta_0,\pi_0}(y)$, $b = f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)$, and $c = \xi/2$. Thus, the first part of Assumption A1(iii) also implies

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\{ \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} = O(\epsilon^{1/2}). \tag{S1}$$

Next, we find

$$\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \mathbb{E}_{\beta_0,\pi_0} q_\epsilon(Y) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} q_\epsilon(Y) \right| \\
&= \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \int_{\mathcal{Y}} q_\epsilon(Y) \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \left[f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right] dy \right| \\
&\leq \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} |q_\epsilon(Y)|^{\frac{\xi}{\xi-1}} \left| \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \right|^{\frac{\xi}{\xi-1}} dy \right\}^{\frac{\xi-1}{\xi}} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} |q_\epsilon(Y)|^{\frac{\xi}{\xi-1}} \left| f_{\beta_0,\pi_0}(y) + f_{\beta_0,\pi(\gamma_*)}(y) \right| dy \right\}^{\frac{\xi-1}{\xi}} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ 2 \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |q_\epsilon(Y)|^{\zeta} \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left| f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&= o(1),
\end{aligned}$$

where the first inequality is an application of Hölder's inequality, the second inequality uses that $\left| \frac{f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y)}{f^{1/\xi}_{\beta_0,\pi_0}(y) - f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)} \right|^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \left[f_{\beta_0,\pi_0}(y) + f_{\beta_0,\pi(\gamma_*)}(y) \right]$,[1] the last line uses that $\kappa = \xi/(\xi-1)$, and the final conclusion follows from our assumptions and (S1).

Part (iii): We have

$$\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) \\
&\qquad - \left\langle \pi_0 - \pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\rangle \\
&= \int_{\mathcal{Y}} h_\epsilon(y, \beta_0, \gamma_*) \left[f_{\beta_0,\pi_0}(y) - f_{\beta_0,\pi(\gamma_*)}(y) - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle f_{\beta_0,\pi(\gamma_*)}(y) \right] dy \\
&= \int_{\mathcal{Y}} h_\epsilon(y, \beta_0, \gamma_*) \left[f^{1/2}_{\beta_0,\pi_0}(y) + f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right]
\end{aligned}$$

[1] For $a, b \geq 0$ there exists $c \in [a, b]$ such that by the mean value theorem we have $(a^\xi - b^\xi)/(a-b) = \xi c^{\xi-1} \leq \xi \max(a^{\xi-1}, b^{\xi-1})$, and therefore $[(a^\xi - b^\xi)/(a-b)]^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \max(a^\xi, b^\xi) \leq \xi^{\xi/(\xi-1)}(a^\xi + b^\xi)$, which we apply here with $a = f^{1/\xi}_{\beta_0,\pi_0}(y)$ and $b = f^{1/\xi}_{\beta_0,\pi(\gamma_*)}(y)$.

$$
\underbrace{\times \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) - \frac{1}{2} \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right] dy}_{=:a^{(1)}_{\beta_0,\gamma_*,\pi_0}}
$$

$$
+ \frac{1}{2} \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\beta_0,\gamma_*) f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right] dy}_{=:a^{(2)}_{\beta_0,\gamma_*,\pi_0}} .
$$

Applying the Cauchy-Schwarz inequality and our assumptions we find that

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| a^{(1)}_{\beta_0,\gamma_*,\pi_0} \right|^2 \\
&\leq 4 \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} h^2_\epsilon(Y,\beta_0,\gamma_*) \right\} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \int_{\mathcal{Y}} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) - \left\langle \pi_0 - \pi(\gamma_*), \nabla_\pi f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right\rangle \right]^2 dy \right\} \\
&= O(\epsilon^{1/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| a^{(2)}_{\beta_0,\gamma_*,\pi_0} \right|^2 \\
&\leq \left\{ \mathbb{E}_{\beta_0,\pi(\gamma_*)} h^2_\epsilon(Y,\beta_0,\gamma_*) \right\} \\
&\qquad \times \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left\| \pi_0 - \pi(\gamma_*) \right\|^2_{\mathrm{ind},\gamma_*} \int_{\mathcal{Y}} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\|^2_{\gamma_*} \left[f^{1/2}_{\beta_0,\pi_0}(y) - f^{1/2}_{\beta_0,\pi(\gamma_*)}(y) \right]^2 dy \right\} \\
&= o(\epsilon).
\end{aligned}
$$

Combining this gives the statement in the lemma. ∎

Proof of Lemma A2. Applying part (ii) of Lemma S1 with $q_\epsilon(y) = h_\epsilon(y,\beta_0,\gamma_*)$ and using the unbiasedness constraint (2) we find that $\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) = o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Part (i) of Lemma S1 guarantees that $\left| \delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right| = o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. We therefore have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right]^2 \\
&\quad = \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) \right]^2 - 2 \left(\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right) \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) + \left(\delta_{\beta_0,\pi_0} - \delta_{\beta_0,\pi(\gamma_*)} \right)^2 \\
&\quad = \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) \right]^2 + o(1),
\end{aligned}
$$

uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. Applying part (ii) of Lemma S1 with $q_\epsilon(y) = [h_\epsilon(y,\beta_0,\gamma_*)]^2$ we find that $\mathbb{E}_{\beta_0,\pi_0} [h_\epsilon(Y,\beta_0,\gamma_*)]^2 = \mathbb{E}_{\beta_0,\pi(\gamma_*)} [h_\epsilon(Y,\beta_0,\gamma_*)]^2 + o(1) = \mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*)) + o(1)$, uniformly in $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where in the last step we have also used that $h_\epsilon(y,\beta_0,\gamma_*)$ satisfies the unbiasedness constraint (2). Therefore,

$$
\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[h_\epsilon(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right]^2 = \mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*)) + o(1). \tag{S2}
$$

Using the unbiasedness constraint again, as well as Lemma S1(iii) and Assumptions A1(ii) and A1(iv) we find

$$
\begin{aligned}
&\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left|\mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right| \\
&= \sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left|\left\langle \pi_0 - \pi(\gamma_*), \mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y,\beta_0,\gamma_*)\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) - \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}\right\rangle\right| + o(\epsilon^{1/2}) \\
&= \epsilon^{1/2} \left\|\mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y,\beta_0,\gamma_*)\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) - \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}\right\|_{\gamma_*} + o(\epsilon^{1/2}) \\
&= b_\epsilon(h_\epsilon,\beta_0,\gamma_*) + o(\epsilon^{1/2}), \qquad \text{(S3)}
\end{aligned}
$$

where in the last step we used the definition of the worst-case bias in (8) of the main text. We furthermore have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[\widehat{\delta}(h_\epsilon,\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right]^2 \\
&= \mathbb{E}_{\beta_0,\pi_0}\left(\frac{1}{n}\sum_{i=1}^n h(Y_i,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right)^2 \\
&= \left[\mathbb{E}_{\beta_0,\pi_0} h(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right]^2 + \frac{1}{n}\mathrm{Var}_{\beta_0,\pi_0}\left[h(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right] \\
&= \frac{n-1}{n}\left[\mathbb{E}_{\beta_0,\pi_0} h(Y,\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0} + \delta_{\beta_0,\pi(\gamma_*)}\right]^2 + \frac{1}{n}\mathbb{E}_{\beta_0,\pi_0}\left[h(Y,\beta_0,\gamma_*) + \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0}\right]^2 .
\end{aligned}
$$

Taking the supremum of this last result over $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, and then applying (S2) and (S3) gives

$$
\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0}\left[\widehat{\delta}(h_\epsilon,\beta_0,\gamma_*) - \delta_{\beta_0,\pi_0}\right]^2 = b_\epsilon(h_\epsilon,\beta_0,\gamma_*)^2 + \frac{\mathrm{Var}_{\beta_0,\pi(\gamma_*)}(h_\epsilon(Y,\beta_0,\gamma_*))}{n} + o(\epsilon),
$$

which is the statement of the lemma. ∎

Proof of Lemma A3. Let $\eta = (\beta', \gamma')'$, $\widehat{\eta} := (\widehat{\beta}', \widehat{\gamma}')'$, and $\eta_* := (\beta_0', \gamma_*')'$. By a Taylor expansion in η around η_* we find that

$$
\begin{aligned}
\widehat{\delta}_\epsilon^{\mathrm{MMSE}} &= \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\widehat{\beta},\widehat{\gamma}) \\
&= \delta_{\beta_0,\pi(\gamma_*)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\beta_0,\gamma_*) \\
&\quad \underbrace{(\widehat{\eta}-\eta_*)'\left[\nabla_\eta \delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\beta_0,\gamma_*)\right]}_{=r^{(1)}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta_*)'\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\beta_0,\gamma_*)\right]}_{=r^{(2)}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta_*)'\left[\mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\beta_0,\gamma_*) - \mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\beta_0,\gamma_*)\right]}_{=r^{(3)}}
\end{aligned}
$$

$$+\frac{1}{2}\underbrace{(\widehat{\eta}-\eta_*)'\left[\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\eta\eta'}h_\epsilon^{\rm MMSE}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right](\widehat{\eta}-\eta_*)}_{=r^{(4)}}, \tag{S4}$$

where $\widetilde{\eta}=(\widetilde{\beta}',\widetilde{\gamma}')'$ is a value between $\widehat{\eta}$ and η_*. Our constraints (2) and (4) guarantee that $\nabla_\eta \delta_{\beta_0,\pi(\gamma_*)}+\mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)=0$; that is, we have $r^{(1)}=0$. Using Assumption A2 and the Cauchy-Schwarz inequality we furthermore find

$$\begin{aligned}
&\left(\mathbb{E}_{\beta_0,\pi_0}\left|r^{(2)}\right|\right)^2\\
&\leq \mathbb{E}_{\beta_0,\pi_0}\left\|\widehat{\eta}-\eta_*\right\|^2\ \mathbb{E}_{\beta_0,\pi_0}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\eta h_\epsilon^{\rm MMSE}(Y_i,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\rm MMSE}(Y_i,\beta_0,\gamma_*)\right]\right\|^2\\
&\leq \mathbb{E}_{\beta_0,\pi_0}\left\|\widehat{\eta}-\eta_*\right\|^2\ \frac{1}{n}\,\mathbb{E}_{\beta_0,\pi_0}\left\|\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)\right\|^2=O\left(\frac{1}{n^2}\right),
\end{aligned}$$

uniformly in $\pi_0\in\Gamma_\epsilon(\gamma_*)$, where in the second step we have used the independence of Y_i across i. Similarly, we have

$$\begin{aligned}
\left(\mathbb{E}_{\beta_0,\pi_0}\left|r^{(3)}\right|\right)^2 &\leq \mathbb{E}_{\beta_0,\pi_0}\left\|\widehat{\eta}-\eta_*\right\|^2\left\|\mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)\right\|^2\\
&=O\left(\frac{1}{n}\right)O\left(\epsilon\right)=O\left(\frac{1}{n^2}\right),
\end{aligned}$$

uniformly in $\pi_0\in\Gamma_\epsilon(\gamma_*)$, where we have used that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\left\|\mathbb{E}_{\beta_0,\pi_0}\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)-\mathbb{E}_{\beta_0,\pi(\gamma_*)}\nabla_\eta h_\epsilon^{\rm MMSE}(Y,\beta_0,\gamma_*)\right\|=O(\epsilon^{1/2}),$$

which follows from Assumptions A1(iii) and A2(ii) by using the proof strategy of part (ii) of Lemma S1. Finally, applying Hölder's inequality we have

$$\begin{aligned}
\mathbb{E}_{\beta_0,\pi_0}\left|r^{(4)}\right| &\leq \mathbb{E}_{\beta_0,\pi_0}\left[\left\|\widehat{\eta}-\eta_*\right\|^2\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\eta\eta'}h_\epsilon^{\rm MMSE}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right\|\right]\\
&\leq \left\{\mathbb{E}_{\beta_0,\pi_0}\left\|\widehat{\eta}-\eta_*\right\|^\chi\right\}^{\frac{2}{\chi}}\left\{\mathbb{E}_{\beta_0,\pi_0}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\eta\eta'}h_\epsilon^{\rm MMSE}(Y_i,\widetilde{\beta},\widetilde{\gamma})\right\|^{\frac{\chi}{\chi-2}}\right\}^{\frac{\chi-2}{\chi}}\\
&=O\left(\frac{1}{n}\right),
\end{aligned}$$

uniformly in $\pi_0\in\Gamma_\epsilon(\gamma_*)$, where we have used Assumption A2(iii). We have thus shown that

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)}\mathbb{E}_{\beta_0,\pi_0}\left|r^{(1)}+r^{(2)}+r^{(3)}+\frac{1}{2}r^{(4)}\right|=O\left(\frac{1}{n}\right),$$

which together with (S4) gives the statement of the lemma. ∎

The proof of the next lemma uses the following theorem of Petrov (1975), which generalizes the Berry-Esseen theorem to sample averages of random variables without a third moment.

Theorem S1 (Theorem 5 on p. 112 in Petrov 1975). *Let $X_1, \ldots, X_n$ be independent random variables, such that $\mathbb{E}X_j = 0$, $\mathbb{E}(X_j^2 g(|X_j|)) < \infty$ for $j = 1, \ldots, n$, and for some function $g : [0, \infty) \to [0, \infty)$ such that both $g(x)$ and $x/g(x)$ are non-decreasing for $x > 0$. We write*

$$\sigma_j^2 = \mathbb{E}X_j^2, \qquad B_n = \sum_{j=1}^{n} \sigma_j^2, \qquad F_n(x) = \Pr\left(B_n^{-1/2} \sum_{j=1}^{n} X_j < x \right).$$

Then there exists an absolute constant $A > 0$ such that

$$\sup_x |F_n(x) - \Phi(x)| \le \frac{A}{B_n g(\sqrt{B_n})} \sum_{j=1}^{n} \mathbb{E}\left(X_j^2 g(X_j) \right).$$

Proof of Lemma A4. # Preliminaries: We first establish some preliminary results on the sample averages of

$$\widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) := h_\epsilon(Y_i, \beta_0, \gamma_*) - \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y_i, \beta_0, \gamma_*).$$

According to our assumptions the $\widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0)$ are independent random variables with zero mean and finite absolute moments of order $\kappa > 2$, under $P_0 = P(\beta_0, \pi_0)$. By applying the result in Dharmadhikari and Jogdeo (1969) we thus find that[2]

$$\mathbb{E}_{\beta_0,\pi_0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right|^\kappa \le C_\kappa \, \mathbb{E}_{\beta_0,\pi_0} \left| \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right|^\kappa,$$

where the constant $C_\kappa > 0$ only depends on κ. Through a combination of the Minkowski and Hölder's inequalities we find that our assumption $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} |h_\epsilon(Y, \beta_0, \gamma_*)|^\kappa = O(1)$ also guarantees $\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left| \widetilde{h}_\epsilon(Y, \beta_0, \gamma_*) \right|^\kappa = O(1)$. We therefore obtain that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right|^\kappa \right)^{\frac{1}{\kappa}} = O(1). \tag{S5}$$

Next, we apply Theorem 5 of Chapter V in Petrov (1975), which is restated above as Theorem S1, with X_i equal to $\widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0)$ and $g(x) = x^{\min\{1, \kappa-2\}}$ to find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \sup_{x \in \mathbb{R}} \left| \mathrm{P}_{\beta_0,\pi_0} \left(\frac{\sum_{i=1}^{n} \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0)}{\sqrt{n}\, \sigma(\beta_0, \gamma_*, \pi_0)} \le x \right) - \Phi(x) \right| = o(1),$$

where $\sigma^2(\beta_0, \gamma_*, \pi_0) = \mathbb{E}_{\beta_0,\pi_0} \widetilde{h}_\epsilon^2(Y_i, \beta_0, \gamma_*, \pi_0)$. This, in particular, implies that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathrm{P}_{\beta_0,\pi_0} \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*, \pi_0) \right| > \log(n) \right) = o(1). \tag{S6}$$

[2] This result is an extension of the Bahr-Esseen inequality to moments larger than two. See also inequality number 16 on p. 60 of Petrov (1975).

By an application of Hölder's inequality we find that (S5) and (S6) also imply

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) \right)^2 \mathbb{1} \left(\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) \right| > \log n \right) \right] = o(1). \tag{S7}$$

Finally, we notice that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left| \delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y, \beta_0, \gamma_*) \right| = O(\epsilon^{1/2}), \tag{S8}$$

which follows by applying part (i) and (ii) of Lemma S1 with $q_\epsilon(y) = h_\epsilon(y, \beta_0, \gamma_*)$ and noting that $\mathbb{E}_{\beta_0,\pi(\gamma_*)} h_\epsilon(Y, \beta_0, \gamma_*) = 0$ by the unbiasedness constraint (2).

Main result of the Lemma A4: Having established those preliminary results, we now derive the statement of the lemma. Define

$$\begin{aligned} k_n &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n h_\epsilon(Y_i, \beta_0, \gamma_*) + \sqrt{n} \left[\delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} \right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) + \sqrt{n} \left[\delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y_i, \beta_0, \gamma_*) \right]. \end{aligned}$$

The decomposition of $\widehat{\delta}_\epsilon$ in (A2) can then be rewritten as

$$\sqrt{n} \left(\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0} \right) = k_n + R_n.$$

We have

$$\begin{aligned} & n\, \mathbb{E}_{\beta_0,\pi_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0} \right)^2 \mathbb{1} \left(\left| \widehat{\delta}_\epsilon - \delta_{\beta_0,\pi_0} \right| \le m_n \right) \right] \\ &\quad = \mathbb{E}_{\beta_0,\pi_0} \left[(k_n + R_n)^2 \, \mathbb{1} \left(|k_n + R_n| \le n^{1/2} m_n \right) \right] \\ &\quad = \mathbb{E}_{\beta_0,\pi_0} k_n^2 - \underbrace{\mathbb{E}_{\beta_0,\pi_0} \left[k_n^2 \, \mathbb{1} \left(|k_n + R_n| > n^{1/2} m_n \right) \right]}_{=\text{term I}} \\ &\qquad + \underbrace{\mathbb{E}_{\beta_0,\pi_0} \left[(R_n^2 + 2 k_n R_n) \mathbb{1} \left(|k_n + R_n| \le n^{1/2} m_n \right) \right]}_{=\text{term II}}. \end{aligned}$$

Thus, Lemma A4 is proved if we can show that term I is $o(1)$, and that term II is larger or equal to minus $o(1)$, both uniformly over $\pi_0 \in \Gamma_\epsilon(\gamma_*)$. For term I we use Hölder's inequality to obtain that

$$\begin{aligned} & \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \left[k_n^2 \, \mathbb{1} \left(|k_n + R_n| > n^{1/2} m_n \right) \right] \\ &\le \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} |k_n|^\kappa \right)^{\frac{2}{\kappa}} \right\} \left\{ \sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left[\mathbb{E}_{\beta_0,\pi_0} \mathbb{1} \left(|k_n + R_n| > n^{1/2} m_n \right) \right]^{\frac{\kappa-2}{\kappa}} \right\} \\ &\le \Bigg\{ \underbrace{\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \left(\mathbb{E}_{\beta_0,\pi_0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_\epsilon(Y_i, \beta_0, \gamma_*) \right|^\kappa \right)^{\frac{2}{\kappa}}}_{=O(1)} \end{aligned}$$

$$
\left. + \underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left(n^{1/2}\left|\delta_{\beta_0,\pi(\gamma_*)} - \delta_{\beta_0,\pi_0} + \mathbb{E}_{\beta_0,\pi_0} h_\epsilon(Y_i,\beta_0,\gamma_*)\right|\right)}_{=O(1)} \right\}
$$

$$
\times \left\{ \Bigg[\underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \mathbb{1}\left(|k_n| > \frac{1}{2} n^{1/2} m_n\right)}_{=o(1)} + \underbrace{\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} \mathbb{1}\left(|R_n| > \frac{1}{2} n^{1/2} m_n\right)}_{=o(1)} \Bigg]^{\frac{\kappa-2}{\kappa}} \right\}
$$

$$
= o(1),
$$

where we also used the definition of k_n together with the triangle inequality, and we employed (S5), (S6) and (S8) and Assumption (ii) of the lemma, together with our assumption that $n^{1/2} m_n \gg \log(n)$ as $n \to \infty$.

Next, for term II we use that $R_n^2 + 2k_n R_n$ is positive whenever $|R_n| > 2|k_n|$ to obtain that

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2 + 2k_n R_n)\mathbb{1}\left(|k_n + R_n| \le n^{1/2} m_n\right)\right] \\
&\quad = \mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2 + 2k_n R_n)\mathbb{1}\left(|k_n + R_n| \le n^{1/2} m_n\right)\mathbb{1}\left(|R_n| \le 2|k_n|\right)\right] \\
&\qquad + \underbrace{\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2 + 2k_n R_n)\mathbb{1}\left(|k_n + R_n| \le n^{1/2} m_n\right)\mathbb{1}\left(|R_n| > 2|k_n|\right)\right]}_{\ge 0} \\
&\quad \ge \mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2 + 2k_n R_n)\mathbb{1}\left(|k_n + R_n| \le n^{1/2} m_n\right)\mathbb{1}\left(|R_n| \le 2|k_n|\right)\right] \\
&\quad \ge -2\,\mathbb{E}_{\beta_0,\pi_0}\Big[|k_n|\,|R_n|\,\mathbb{1}\left(|R_n| \le 2|k_n|\right)\Big] \\
&\quad \ge -2\left\{\mathbb{E}_{\beta_0,\pi_0} k_n^2\right\}^{1/2}\left\{\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n| \le 2|k_n|\right)\right]\right\}^{1/2}
\end{aligned}
$$

where in the last step we also used the Cauchy-Schwarz inequality. Our preliminary results (S5) and (S8) imply that $\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0,\pi_0} k_n^2 = O(1)$. Furthermore we have

$$
\begin{aligned}
&\mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n| \le 2|k_n|\right)\right] \\
&\quad = \mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n| \le 2|k_n|\right)\mathbb{1}\left(|k_n| \le \log n\right)\right] \\
&\qquad + \mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n| \le 2|k_n|\right)\mathbb{1}\left(|k_n| > \log n\right)\right] \\
&\quad \le \mathbb{E}_{\beta_0,\pi_0}\left[R_n^2\,\mathbb{1}\left(|R_n| \le 2\,\log n\right)\right] + 4\,\mathbb{E}_{\beta_0,\pi_0}\left[k_n^2\,\mathbb{1}\left(|k_n| > \log n\right)\right] \\
&\quad = o(1),
\end{aligned}
$$

uniformly over $\pi_0 \in \Gamma_\epsilon(\gamma_*)$, where we used (S7) and Assumption (v) of the lemma. We thus conclude that term II indeed satisfies

$$
\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left\{-\mathbb{E}_{\beta_0,\pi_0}\left[(R_n^2 + 2k_n R_n)\mathbb{1}\left(|k_n + R_n| \le n^{1/2} m_n\right)\right]\right\} \le o(1).
$$

Combining the above gives the statement of the lemma. ∎

S1.2 Lemma A1

Notation. For the proof of Lemma A1 (which assumes the locally quadratic case of Section 3) it is convenient to introduce some further notation. We assume that there exists a map

$\Omega_{\gamma_*} : \overline{\mathcal{T}} \to \mathcal{T}$ such that, for all $v \in \overline{\mathcal{T}}$,

$$\|v\|^2_{\text{ind},\gamma_*} = \left\langle v, \Omega_{\gamma_*} v \right\rangle .$$

We assume that Ω_{γ_*} is invertible, and write $\Omega^{-1}_{\gamma_*} : \mathcal{T} \to \overline{\mathcal{T}}$ for its inverse. The map $\Omega^{-1}_{\gamma_*}$ is exactly the "transposition" map introduced less formally in the main text; that is, for $u \in \mathcal{T}$ we have $u^\top = \Omega^{-1}_{\gamma_*} u \in \overline{\mathcal{T}}$. Thus, our norm on the cotangent space from the main text $\|u\|^2_{\gamma_*} = u^\top u$ can now be written as

$$\|u\|^2_{\gamma_*} = \left\langle \Omega^{-1}_{\gamma_*} u, u \right\rangle .$$

The norm $\|\cdot\|_{\gamma_*}$ is dual to $\|\cdot\|_{\text{ind},\gamma_*}$; that is, we have

$$\|u\|_{\gamma_*} = \sup_{v \in \overline{\mathcal{T}} \setminus \{0\}} \frac{\langle v, u \rangle}{\|v\|_{\text{ind},\gamma_*}}.$$

Notice also that $\|\cdot\|_{\text{ind},\gamma_*}$, $\|\cdot\|_{\gamma_*}$, Ω_{γ_*}, and $\Omega^{-1}_{\gamma_*}$ could all be defined for general $\pi \in \Pi$, but since we use them only at the reference value $\pi(\gamma_*)$ we index them simply by γ_*.

The vector norms $\|\cdot\|_{\text{ind},\gamma_*}$, $\|\cdot\|_{\gamma_*}$ and $\|.\|$ on $\overline{\mathcal{T}}$, $\mathcal{T}$ and $\mathbb{R}^{\dim\beta+\dim\gamma}$ induce natural norms on any maps between $\overline{\mathcal{T}}$, $\mathcal{T}$ and $\mathbb{R}^{\dim\beta+\dim\gamma}$. With a slight abuse of notation we denote all those norms simply by $\|.\|_{\gamma_*}$. In particular, for $\Omega^{-1}_{\gamma_*} : \mathcal{T} \to \overline{\mathcal{T}}$ we have

$$\left\|\Omega^{-1}_{\gamma_*}\right\|_{\gamma_*} := \sup_{u \in \mathcal{T} \setminus \{0\}} \frac{\|\Omega^{-1}_{\gamma_*} u\|_{\text{ind},\gamma_*}}{\|u\|_{\gamma_*}} = \sup_{u \in \mathcal{T} \setminus \{0\}} \frac{\left\langle \Omega^{-1}_{\gamma_*} u, , u \right\rangle^{1/2}}{\|u\|_{\gamma_*}} = 1, \tag{S9}$$

and for $H_{\pi,\beta\gamma} : \mathbb{R}^{\dim\beta+\dim\gamma} \to \mathcal{T}$ defined in Section 3.1 we have

$$\|H_{\pi,\beta\gamma}\|_{\gamma_*} := \sup_{w \in \mathbb{R}^{\dim\beta+\dim\gamma} \setminus \{0\}} \frac{\|H_{\pi,\beta\gamma} w\|_{\gamma_*}}{\|w\|} = \sup_{v \in \overline{\mathcal{T}} \setminus \{0\}} \sup_{w \in \mathbb{R}^{\dim\beta+\dim\gamma} \setminus \{0\}} \frac{\langle v, H_{\pi,\beta\gamma} w \rangle}{\|v\|_{\text{ind},\gamma_*} \|w\|}.$$

Using Assumption A1(v) and the Cauchy-Schwarz inequality we find that

$$\begin{aligned}
\|H_{\pi,\beta\gamma}\|_{\gamma_*} &= \left\| \mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\{ \left[\nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y)\right] \left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y)\right]' \right\} \right\|_{\gamma_*} \\
&\le \left[\mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\| \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|^2_{\gamma_*} \right]^{1/2} \left[\mathbb{E}_{\beta_0,\pi(\gamma_*)} \left\| \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y) \right\|^2 \right]^{1/2} \\
&= O(1).
\end{aligned} \tag{S10}$$

Proof of Lemma A1. Equation (20) in Lemma 1 in the main text provides an explicit solution for $h^{\text{MMSE}}_\epsilon(y, \beta_0, \gamma_*)$, which in the notation of this appendix can be written as

$$\begin{aligned}
h^{\text{MMSE}}_\epsilon(y, \beta_0, \gamma_*) &= \left[\nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}\right]' H^{-1}_{\beta\gamma} \left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)\right] \\
&\quad + \left\langle \left[\widetilde{H}_\pi \, \Omega_{\gamma_*} + (\epsilon n)^{-1} \Omega_{\gamma_*} \right]^{-1} \widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} \, , \, \widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) \right\rangle ,
\end{aligned}$$

where $\widetilde{\nabla}_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) = \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(y) - H_{\pi,\beta\gamma} H^{-1}_{\beta\gamma} \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)$ and $\widetilde{\nabla}_\pi \delta_{\beta_0,\pi(\gamma_*)} = \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)} - H_{\pi,\beta\gamma} H^{-1}_{\beta\gamma} \nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}$. We thus have

$$\left|h^{\text{MMSE}}_\epsilon(y, \beta_0, \gamma_*)\right| \le \left\|\nabla_{\beta\gamma} \delta_{\beta_0,\pi(\gamma_*)}\right\| \left\|H^{-1}_{\beta\gamma}\right\| \left\|\left[\nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(y)\right]\right\|$$

$$+ (\epsilon n) \left\| \Omega_{\gamma_*}^{-1} \right\|_{\gamma_*} \left\| \widetilde{\nabla}_\pi \delta_{\beta_0, \pi(\gamma_*)} \right\|_{\gamma_*} \left\| \widetilde{\nabla}_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|_{\gamma_*},$$

where we used that $\left\| \left[\widetilde{H}_\pi \Omega_{\gamma_*} + (\epsilon n)^{-1} \Omega_{\gamma_*} \right]^{-1} \right\|_{\gamma_*} \leq (\epsilon n) \left\| \Omega_{\gamma_*}^{-1} \right\|_{\gamma_*}$, because both $\widetilde{H}_\pi \Omega_{\gamma_*}$ and Ω_{γ_*} are positive semi-definite. We furthermore have

$$\begin{aligned} \left\| \widetilde{\nabla}_\pi \delta_{\beta_0, \pi(\gamma_*)} \right\|_{\gamma_*} &\leq \left\| \nabla_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|_{\gamma_*} + \left\| H_{\pi, \beta\gamma} \right\|_{\gamma_*} \left\| H_{\beta\gamma}^{-1} \right\| \left\| \nabla_{\beta\gamma} \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|, \\ \left\| \widetilde{\nabla}_\pi \log f_{\beta_0, \pi(\gamma_*)}(y) \right\|_{\gamma_*} &\leq \left\| \nabla_\pi \delta_{\beta_0, \pi(\gamma_*)} \right\|_{\gamma_*} + \left\| H_{\pi, \beta\gamma} \right\|_{\gamma_*} \left\| H_{\beta\gamma}^{-1} \right\| \left\| \nabla_{\beta\gamma} \delta_{\beta_0, \pi(\gamma_*)} \right\|. \end{aligned}$$

Combining those inequalities with our Assumption A1(ii) and (v) as well as the results (S9) and (S10) above we find that

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0} \left[h_\epsilon^{\mathrm{MMSE}}(Y, \beta_0, \gamma_*) \right]^{2+\nu} = O(1).$$

∎

S1.3 Lemma 1

Before deriving the equivalent characterizations of $h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$ given in the lemma we note that the optimization problem (10) that defines $h_\epsilon^{\mathrm{MMSE}}(y, \beta_0, \gamma_*)$ has a unique solution (up to possible deviations on a measure zero set of y's, which are irrelevant for our purposes). This uniqueness follows, because under the unbiasedness constraint (2), we have $\mathrm{Var}_{\beta_0, \pi(\gamma_*)}(h(Y, \beta_0, \gamma_*)) = \mathbb{E}_{\beta_0, \pi(\gamma_*)} h^2(Y, \beta_0, \gamma_*)$, which is quadratic and strictly convex in $h(y, \beta_0, \gamma_*)$, while all other components of the objective function and constraints in (10) are linear in $h(y, \beta_0, \gamma_*)$.

Equation (18). Using simplified notation here, our goal is to find the function $h(y) = h(y, \beta_0, \gamma_*)$ that minimizes

$$\mathbb{E} h^2(Y) + (\epsilon n) \left\{ \nabla_\pi \delta - \mathbb{E}\left[h(Y) s_\pi(Y) \right] \right\}^\top \left\{ \nabla_\pi \delta - \mathbb{E}\left[h(Y) s_\pi(Y) \right] \right\},$$

subject to the constraints $\mathbb{E} h(Y) = 0$ and $\mathbb{E}\, h(Y) s_{\beta\gamma}(Y) = \nabla_{\beta\gamma} \delta$.

Using the latter constraint and the definition of $\widetilde{\nabla}_\pi$ we can equivalently rewrite the objective function as

$$\begin{aligned} \mathbb{E} h^2(Y) + (\epsilon n) \left\{ \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h(Y) \widetilde{s}_\pi(Y) \right] \right\}^\top \left\{ \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h(Y) \widetilde{s}_\pi(Y) \right] \right\} \\ + 2 \left\{ \nabla_{\beta\gamma} \delta - \mathbb{E}\left[h(Y) s_{\beta\gamma}(Y) \right] \right\}' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta. \end{aligned}$$

The unconstrained minimizer of this rewritten quadratic objective function satisfies the first-order condition

$$h_\epsilon^{\mathrm{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma} \delta + (\epsilon n)\, \widetilde{s}_\pi(y)^\top \left\{ \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y) \widetilde{s}_\pi(Y) \right] \right\},$$

and because $\mathbb{E} s_{\beta\gamma}(Y) = 0$, $\mathbb{E} \widetilde{s}_\pi(Y) = 0$, and $\mathbb{E}[s_{\beta\gamma}(Y) s_{\beta\gamma}(Y)'] = H_{\beta\gamma}$, we find that this unconstrained minimizer already satisfies both constraints $\mathbb{E} h(Y) = 0$ and $\mathbb{E} h(Y) s_{\beta\gamma}(Y) = \nabla_{\beta\gamma} \delta$, and is therefore also the constrained minimizer that we wanted to derive.

Equation (19). Note that, by (18), we have $h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}\delta + \widetilde{s}_\pi(y)^\top u$, for some $u \in \mathcal{T}$, and one can easily verify that this implies that $\widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y)\widetilde{s}_\pi(Y)\right]$ is equal to the same expression with $\widetilde{s}_\pi$ replaced by s_π.

Equation (20). We have already shown that equation (18) is the FOC of the minimization problem (10). We now want to show that the solution for $h_\epsilon^{\text{MMSE}}(y)$ given in equation (20) satisfies the FOC (18), which implies that it solves (10). Equation (18) can be rewritten as

$$h_\epsilon^{\text{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \nabla_{\beta\gamma}\delta + (\epsilon n)\, \widetilde{s}_\pi(y)^\top u, \quad u := \widetilde{\nabla}_\pi \delta - \mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y)\widetilde{s}_\pi(Y)\right]. \tag{S11}$$

Plugging the expression for $h_\epsilon^{\text{MMSE}}(y)$ given by equation (20) into this definition of u and using that $\mathbb{E}\left[\widetilde{s}_\pi(Y)\widetilde{s}_\pi(Y)^\top\right] = \widetilde{H}_\pi$, and $\mathbb{E}\left[\widetilde{s}_\pi(Y) s_{\beta\gamma}(Y)'\right] = 0$, we find that (20) implies that

$$\begin{aligned}
u &= \widetilde{\nabla}_\pi \delta - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]^{-1} \widetilde{\nabla}_\pi \delta \\
&= \left\{\mathbb{I} - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]^{-1}\right\} \widetilde{\nabla}_\pi \delta \\
&= \left\{\left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]\left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]^{-1} - \widetilde{H}_\pi \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]^{-1}\right\} \widetilde{\nabla}_\pi \delta \\
&= (\epsilon n)^{-1} \left[\widetilde{H}_\pi + (\epsilon n)^{-1}\mathbb{I}\right]^{-1} \widetilde{\nabla}_\pi \delta.
\end{aligned}$$

This expression for u makes the first equation in (S11) equivalent to (20). Therefore, we have shown that $h_\epsilon^{\text{MMSE}}(y)$ as given by (20) indeed solves (18), and therefore also our optimization problem in (10).

S1.4 Lemma 2

Our goal is to choose the function $h(\cdot, \cdot, \beta, \gamma, f_X)$ such that the worst-case mean squared error

$$\sup_{\pi_0 \in \Gamma_\epsilon(\gamma_*)} \mathbb{E}_{\beta_0, \pi_0, f_X}\left[\left(\widehat{\delta}_h - \delta_{\beta_0,\pi_0,f_X}\right)^2\right]$$

is minimized for small values of ϵ, subject to unbiasedness under the reference model, and also subject to local robustness constraints to account for the fact that β_0, γ_* and f_X are estimated from the sample.

Unbiasedness is

$$\mathbb{E}_{f_X}\, \mathbb{E}_{\beta_0, \pi(\gamma_*)}\, h(Y, X, \beta_0, \gamma_*, f_X) = 0, \tag{S12}$$

while local robustness is

$$\begin{aligned}
&\mathbb{E}_{f_X}\, \mathbb{E}_{\beta_0,\pi(\gamma_*)}\, h(Y, X, \beta_0, \gamma_*, f_X)\, \nabla_{\beta\gamma} \log f_{\beta_0,\pi(\gamma_*)}(Y \mid X) = \mathbb{E}_{f_X}\, \nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)}(X), \\
&\mathbb{E}_{\beta_0,\pi(\gamma_*)}\, \left[h(Y, X, \beta_0, \gamma_*, f_X) \mid X = x\right] = \delta_{\beta_0,\pi(\gamma_*)}(x) - \mathbb{E}_{f_X}\, \delta_{\beta_0,\pi(\gamma_*)}(X).
\end{aligned} \tag{S13}$$

The minimum-MSE influence function satisfies

$h_\epsilon^{\text{MMSE}}(\cdot, \cdot, \beta_0, \gamma_*, f_X) =$

$$\underset{h(\cdot,\cdot,\beta_0,\gamma_*,f_X)}{\operatorname{argmin}} \left\{ \epsilon \left\| \mathbb{E}_{f_X} \nabla_\pi \delta_{\beta_0,\pi(\gamma_*)}(X) - \mathbb{E}_{f_X} \mathbb{E}_{\beta_0,\pi(\gamma_*)} h(Y,X,\beta_0,\gamma_*,f_X) \nabla_\pi \log f_{\beta_0,\pi(\gamma_*)}(Y \mid X) \right\|_{\gamma_*}^2 + \frac{\mathbb{E}_{f_X} \operatorname{Var}_{\beta_0,\pi(\gamma_*)}(h(Y,X,\beta_0,\gamma_*,f_X) \mid X)}{n} \right\} \quad \text{subject to (S12) and (S13).}$$

In the locally quadratic case, following similar derivations as for equation (18) in Lemma 1, we obtain (21).

S1.5 Corollary 1

This is a direct implication of (20).

S1.6 Corollary 2

This is a direct implication of (19).

S1.7 Corollary 3

Lemma 2 implies, analogously to (19), that

$$\begin{aligned} h_\epsilon^{\text{MMSE}}(y,x) \;=\; & \delta(x) - \mathbb{E}_{f_X}\delta(X) + s_{\beta\gamma}(y \mid x)'[\mathbb{E}_{f_X} H_{\beta\gamma}(X)]^{-1}\, \mathbb{E}_{f_X} \nabla_{\beta\gamma}\delta(X) \\ & + (\epsilon n)\widetilde{s}_\pi(y \mid x)^\top \left\{ \mathbb{E}_{f_X} \nabla_\pi \delta(X) - \mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y,X) s_\pi(Y \mid X)\right] \right\}. \end{aligned} \tag{S14}$$

Since A and X are independent, $\mathbb{E}_{f_X} \nabla_\pi \delta(X)$ can be represented by the function

$$a \mapsto \mathbb{E}_{f_X}\left[\Delta(a,X)\right] - \mathbb{E}_{f_X}\delta(X).$$

Likewise, $\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y,X) s_\pi(Y \mid X)\right]$ can be represented by the function

$$a \mapsto \mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y,X) \mid A = a, X\right] = \overline{h}_\epsilon^{\text{MMSE}}(a).$$

Moreover, we have for any cotangent element u (a function of a),

$$\begin{aligned} \widetilde{s}_\pi(y \mid x)^\top u =& \mathbb{E}\left[u(A) \,\middle|\, Y = y, X = x\right] - \mathbb{E}\left[u(A)\right] \\ & - s_{\beta\gamma}(y \mid x)'[\mathbb{E}_{f_X} H_{\beta\gamma}(X)]^{-1}\mathbb{E}_{f_X}\mathbb{E}\left[s_{\beta\gamma}(Y \mid X) u(A)\right]. \end{aligned} \tag{S15}$$

Corollary 3 then follows from evaluating (S15) at

$$u(a) := \mathbb{E}_{f_X}\left[\Delta(a,X)\right] - \mathbb{E}_{f_X}\delta(X) - \overline{h}_\epsilon^{\text{MMSE}}(a).$$

S1.8 Corollary 4

Let us start again from (S14). In the correlated case, $\mathbb{E}_{f_X} \nabla_\pi \delta(X)$ can be represented by the function

$$(a,x) \mapsto \Delta(a,x) f_X(x) - \delta(x) f_X(x).$$

Likewise, $\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)s_\pi(Y\,|\,X)\right]$ can be represented by the function

$$\begin{aligned}(a,x)\mapsto&\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)\,|\,A=a,X=x\right]f_X(x)-\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)\,|\,X=x\right]f_X(x)\\&=\overline{h}_\epsilon^{\mathrm{MMSE}}(a,x)f_X(x)-\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)\,|\,X=x\right]f_X(x).\end{aligned}$$

Now, by (S13) we have

$$\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)\,|\,X=x\right]=\delta(x)-\mathbb{E}_{f_X}\delta(X). \tag{S16}$$

Hence, $\mathbb{E}_{f_X}\nabla_\pi\delta(X)-\mathbb{E}_{f_X}\mathbb{E}\left[h_\epsilon^{\mathrm{MMSE}}(Y,X)s_\pi(Y\,|\,X)\right]$ can be represented by the function

$$(a,x)\mapsto\Delta(a,x)f_X(x)-\mathbb{E}_{f_X}\delta(X)f_X(x)-\overline{h}_\epsilon^{\mathrm{MMSE}}(a,x)f_X(x).$$

In the present case, cotangent elements are functions of a and x. The corresponding squared dual norm is[3]

$$\|u\|_{\gamma_*}^2=\mathbb{E}_{f_X}\mathbb{E}\left[\left(\frac{u(A,X)-\mathbb{E}[u(A,X)\,|\,X]}{f_X(X)}\right)^2\right].$$

In addition we have, for any cotangent element u (a function of a and x)

$$\begin{aligned}\widetilde{s}_\pi(y\,|\,x)^\top u=&\mathbb{E}\left[\frac{u(A,X)}{f_X(X)}\,|\,Y=y,X=x\right]-\mathbb{E}\left[\frac{u(A,X)}{f_X(X)}\,|\,X=x\right]\\&-s_{\beta\gamma}(y\,|\,x)'[\mathbb{E}_{f_X}H_{\beta\gamma}(X)]^{-1}\mathbb{E}_{f_X}\mathbb{E}\left[s_{\beta\gamma}(Y\,|\,X)\frac{u(A,X)}{f_X(X)}\right].\end{aligned} \tag{S17}$$

Corollary 4 then follows from evaluating (S17) at

$$u(a,x):=\Delta(a,x)f_X(x)-\mathbb{E}_{f_X}\delta(X)f_X(x)-\overline{h}_\epsilon^{\mathrm{MMSE}}(a,x)f_X(x),$$

and noting that, by (S16), $\mathbb{E}[u(A,X)\,|\,X=x]=0$.

S2 Complements to Section 3

S2.1 Dual of the Kullback-Leibler divergence

Let A be a random variable with domain $\mathcal{A}$, reference distribution $f_*(a)$ and "true" distribution $f_0(a)$. We use notation $f_*(a)$ and $f_0(a)$ as if those were densities, but point masses are also allowed. Twice the Kullback-Leibler (KL) divergence reads

$$d(f_0,f_*)=-2\,\mathbb{E}_0\log\frac{f_*(A)}{f_0(A)},$$

[3] This can be shown as in Subsection S2.1, with the difference that here twice the KL divergence reads, using the notation of that subsection, $d(f_0,f_*)=-2\,\mathbb{E}_{f_X}\mathbb{E}_0\log\frac{f_*(A\,|\,X)}{f_0(A\,|\,X)}$. Alternatively, Corollary 4 can be derived by defining π_0 as the joint distribution of (A,X), and imposing the constraint that $\int_{\mathcal{A}}\pi_0(a,x)da=f_X(x)$.

where $\mathbb{E}_0$ is the expectation under f_0. Let $\mathcal{F}$ be the set of all distributions, in particular, $f \in \mathcal{F}$ implies $\int_{\mathcal{A}} f(a)da = 1$. Let $q : \mathcal{A} \to \mathbb{R}$ be a real valued function. For given $f_* \in \mathcal{F}$ and $\epsilon > 0$ we define

$$\|q\|_{*,\epsilon} := \max_{\{f_0 \in \mathcal{F} \,:\, d(f_0, f_*) \leq \epsilon\}} \frac{\mathbb{E}_0\, q(A) - \mathbb{E}_*\, q(A)}{\sqrt{\epsilon}},$$

where $\mathbb{E}_*$ is the expectation under f_*.

We have the following result.

Lemma S2. *For $q : \mathcal{A} \to \mathbb{R}$ and $f_* \in \mathcal{F}$ we assume that the moment-generating function $m_*(t) = \mathbb{E}_* \exp(t\, q(A))$ exists for $t \in (\delta_-, \delta_+)$ and some $\delta_- < 0$ and $\delta_+ > 0$.*[4] *For $\epsilon \in (0, \delta_+^2)$ we then have*

$$\|q\|_{*,\epsilon} = \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}).$$

Proof. Let the cumulant-generating function of the random variable $q(A)$ under the reference measure f_* be $k_*(t) = \log m_*(t)$. We assume existence of $m_*(t)$ and $k_*(t)$ for $t \in (\delta_-, \delta_+)$. This also implies that all derivatives of $m_*(t)$ and $k_*(t)$ exist in this interval. We denote the p-th derivative of $m_*(t)$ by $m_*^{(p)}(t)$, and analogously for $k_*(t)$.

In the following we denote the maximizing f_0 in the definition of $\|q\|_{*,\epsilon}$ simply by f_0. Applying standard optimization method (Karush-Kuhn-Tucker) we find the well-known exponential tilting result

$$f_0(a) = c\, f_*(a)\, \exp(t\, q(a)),$$

where the constants $c, t \in (0, \infty)$ are determined by the constraints $\int_{\mathcal{A}} f_0(a)da = 1$ and $d(f_0, f_*) = \epsilon$. Using the constraint $\int_{\mathcal{A}} f_0(a)da = 1$ we can solve for c to obtain

$$f_0(a) = \frac{f_*(a)\, \exp(t\, q(a))}{\mathbb{E}_* \exp(t\, q(A))} = \frac{f_*(a)\, \exp(t\, q(a))}{m_*(t)}.$$

Using this we find that

$$\begin{aligned}
d(t) &:= d(f_0, f_*) \\
&= 2\, \mathbb{E}_* \frac{f_0(A)}{f_*(A)} \log \frac{f_0(A)}{f_*(A)} \\
&= \frac{2\, t}{m_*(t)}\, \mathbb{E}_* \exp(t\, q(A)) q(A) - \frac{2 \log m_*(t)}{m_*(t)}\, \mathbb{E}_* \exp(t\, q(A)) \\
&= \frac{2\, t\, m_*^{(1)}(t)}{m_*(t)} - 2 \log m_*(t) \\
&= 2 \left[t\, k_*^{(1)}(t) - k_*(t) \right].
\end{aligned}$$

We have $d(0) = 0$, $d^{(1)}(0) = 0$, $d^{(2)}(0) = 2k_*^{(2)}(0) = 2\mathrm{Var}_*(q(A))$, $d^{(3)}(t) = 4k_*^{(3)}(t) + 2tk_*^{(4)}(t)$. A mean-value expansion thus gives

$$d(t) = \mathrm{Var}_*(q(A)) t^2 + \frac{t^3}{6} \left[4\, k_*^{(3)}(\tilde{t}) + 2\, \tilde{t}\, k_*^{(4)}(\tilde{t}) \right],$$

[4] Existence of $m_*(t)$ in an open interval around zero is equivalent to having an exponential decay of the tails of the distribution of the random variable $Q = q(A)$. If $q(a)$ is bounded, then $m_*(t)$ exists for all $t \in \mathbb{R}$.

where $0 \leq \tilde{t} \leq t \leq \delta_+$. The value t that satisfies the constraint $d(t) = \epsilon$ therefore satisfies

$$t = \frac{\epsilon^{\frac{1}{2}}}{\sqrt{\mathrm{Var}_*(q(A))}} + O(\epsilon).$$

Next, using that $\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}}\, \mathbb{E}_* \left[\left(\frac{f_0(A)}{f_*(A)} - 1 \right) q(A) \right]$ we find

$$\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}} \left[k_*^{(1)}(t) - k_*^{(1)}(0) \right].$$

Again using that $k_*^{(2)}(0) = \mathrm{Var}_*(q(A))$ and applying a mean value expansion we obtain

$$\begin{aligned} \|q\|_{*,\epsilon} &= \epsilon^{-\frac{1}{2}} \left[t\, k_*^{(2)}(t) + \frac{1}{2}\, t^2\, k_*^{(3)}(\bar{t}) \right] \\ &= \epsilon^{-\frac{1}{2}} \left[t\, \mathrm{Var}_*(q(A)) + \frac{1}{2}\, t^2\, k_*^{(3)}(\bar{t}) \right] \\ &= \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}), \end{aligned}$$

where $\bar{t} \in [0, t]$. ∎

S2.2 Equations (25), (26) and (27)

Here we use simplified notation as in Section 3. Let us start by deriving (25). In this case β_0 and γ_* are known, and Corollary 2 gives

$$h_\epsilon^{\mathrm{MMSE}} = (\epsilon n) \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \left[\Delta - \delta - \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} h^{\mathrm{MMSE}} \right],$$

so

$$h_\epsilon^{\mathrm{MMSE}} = \left[(\epsilon n)^{-1} \mathbb{I}_{\mathcal{Y}} + \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \right]^{-1} \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \left[\Delta - \delta \right].$$

(25) then follows from the operator identity

$$\left[(\epsilon n)^{-1} \mathbb{I}_{\mathcal{Y}} + \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \right]^{-1} \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} = \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \left[\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} + (\epsilon n)^{-1} \mathbb{I}_{\mathcal{A}} \right]^{-1}.$$

Let us now derive (26). In this case γ_* is known. Since $\Delta(A) = c'\beta_0 = \delta$, Corollary 2 implies

$$h_\epsilon^{\mathrm{MMSE}}(y) = s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} c - (\epsilon n) \left\{ \mathbb{E} \left[\overline{h}^{\mathrm{MMSE}}(A) \,|\, Y = y \right] - s_{\beta\gamma}(y)' H_{\beta\gamma}^{-1} \mathbb{E} \left[s_{\beta\gamma}(Y) \overline{h}^{\mathrm{MMSE}}(A) \right] \right\}.$$

Hence, we have, for some vector b,

$$h_\epsilon^{\mathrm{MMSE}} = s_{\beta\gamma}(y)' b - (\epsilon n) \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} h^{\mathrm{MMSE}}.$$

Using the Woodbury identity

$$\left[\mathbb{I}_{\mathcal{Y}} + (\epsilon n) \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \circ \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \right]^{-1} = \underbrace{\mathbb{I}_{\mathcal{Y}} - \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} \left[\mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y}} + (\epsilon n)^{-1} \mathbb{I}_{\mathcal{A}} \right]^{-1} \mathbb{E}_{\mathcal{Y}\,|\,\mathcal{A}}}_{= \mathbb{W}^\epsilon},$$

we thus obtain

$$h_\epsilon^{\text{MMSE}} = \mathbb{W}^\epsilon s_{\beta\gamma}(y)'b.$$

Lastly, since by (4) $\mathbb{E}\left[h_\epsilon^{\text{MMSE}}(Y)s_{\beta\gamma}(Y)\right] = c$, we obtain (26) whenever the denominator is non-singular.

Finally, let us derive (27). In this case β_0 and γ_* are known and $\Delta(A)$ does not depend on X, and Corollary 3 gives

$$h_\epsilon^{\text{MMSE}} = (\epsilon n)\mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}}\left[\mathbb{E}_{f_X}(\Delta - \delta) - \mathbb{E}_{\mathcal{Y},\mathcal{X}\,|\,\mathcal{A}} h^{\text{MMSE}}\right].$$

Hence, denoting $\mathbb{I}_{\mathcal{Y},\mathcal{X}}h(y,x) = h(y,x)$ the identity operator, we have

$$h_\epsilon^{\text{MMSE}} = \left[(\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y},\mathcal{X}} + \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}} \circ \mathbb{E}_{\mathcal{Y},\mathcal{X}\,|\,\mathcal{A}}\right]^{-1} \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}}\mathbb{E}_{f_X}(\Delta - \delta).$$

(27) then follows from

$$\left[(\epsilon n)^{-1}\mathbb{I}_{\mathcal{Y},\mathcal{X}} + \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}} \circ \mathbb{E}_{\mathcal{Y},\mathcal{X}\,|\,\mathcal{A}}\right]^{-1} \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}} = \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}}\left[\mathbb{E}_{\mathcal{Y},\mathcal{X}\,|\,\mathcal{A}} \circ \mathbb{E}_{\mathcal{A}\,|\,\mathcal{Y},\mathcal{X}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1}.$$

S3 Computation in semi-parametric mixture models

Here we describe how we compute a numerical approximation to the minimum-MSE estimator in semi-parametric mixture models

$$\widehat{\delta}_\epsilon^{\text{MMSE}} = \mathbb{E}_{\widehat{\beta},\pi(\widehat{\gamma})}\,\Delta_{\widehat{\beta}}(A) + \frac{1}{n}\sum_{i=1}^{n} h_\epsilon^{\text{MMSE}}(Y_i, \widehat{\beta}, \widehat{\gamma}),$$

where h_ϵ^{MMSE} is given by Corollary 2, and $\widehat{\beta}, \widehat{\gamma}$ are preliminary estimates. As we pointed out in Section 3, h_ϵ^{MMSE} is the solution to a (well-posed) Tikhonov-regularized linear inverse problem, and many numerical methods are available to solve such problems; see Engl *et al.* (2000) and Kress (2014) for classic references. The simulation-based approach that we have implemented and describe here is closely related to the strategy presented in Bonhomme (2012). We abstract from conditioning covariates. In the presence of correlated covariates X_i, we use the same technique to approximate $h_\epsilon^{\text{MMSE}}(\cdot\,|\,x)$ for each value of $X_i = x$. We use this approach in the numerical illustration based on the dynamic panel data model in Section 6, where the covariate is the initial condition. We denote $\eta = (\beta', \gamma')'$.[5]

Draw an i.i.d. sample $(Y^{(1)}, A^{(1)}), ..., (Y^{(S)}, A^{(S)})$ of S draws from $g_\beta \times \pi(\gamma)$. Let G be $S \times S$ with (τ, s) element $g_\beta(Y^{(\tau)}\,|\,A^{(s)})/\sum_{s'=1}^{S} g_\beta(Y^{(\tau)}\,|\,A^{(s')})$, G_Y be $N \times S$ with (i,s) element $g_\beta(Y_i\,|\,A^{(s)})/\sum_{s'=1}^{S} g_\beta(Y_i\,|\,A^{(s')})$, Δ be $S \times 1$ with s-th element $\Delta_\beta(A^{(s)})$, I be the $S \times S$ identity matrix, and ι and ι_Y be the $S \times 1$ and $N \times 1$ vectors of ones. In addition, let D be the $S \times \dim\eta$ matrix with (s,k) element

$$d_{\eta_k}(Y^{(s)}) = \frac{\sum_{s'=1}^{S}\left(\nabla_{\eta_k}\log g_\beta(Y^{(s)}\,|\,A^{(s')}) + \nabla_{\eta_k}\log\pi(\gamma)(A^{(s')})\right) g_\beta(Y^{(s)}\,|\,A^{(s')})}{\sum_{s'=1}^{S} g_\beta(Y^{(s)}\,|\,A^{(s')})},$$

[5]Here we present a general method based on simulations. In the cross-sectional probit model (30), explicit closed-form expressions are available, and we use those for computation in our first illustration.

and let D_Y be $N \times \dim \eta$ with (i,k) element $d_{\eta_k}(Y_i)$, $Q = I - DD^\dagger$, $\widetilde{G}_Y = G_Y - D_Y D^\dagger G$, $\widetilde{\iota}_Y = \iota_Y - D_Y D^\dagger \iota$, $\widetilde{G} = QG$, $\widetilde{\iota} = Q\iota$, and $\partial\Delta$ be the $K \times 1$ vector with k-th element $\frac{1}{S}\sum_{s=1}^{S} \nabla_{\eta_k}\Delta(A^{(s)},\beta) + \Delta(A^{(s)},\beta)\nabla_{\eta_k}\log\pi(\gamma)(A^{(s)})$.

From Corollary 2, a fixed-S approximation to the minimum-MSE estimator is then

$$\widetilde{\delta}_\epsilon^{\text{MMSE}} = \iota^\dagger \Delta + \iota_Y^\dagger \widetilde{h}_\epsilon^{\text{MMSE}},$$

where

$$\widetilde{h}_\epsilon^{\text{MMSE}} = D_Y (D'D/S)^{-1}\,\partial\Delta + (\epsilon n)\Bigg[\left(\widetilde{G}_Y - \widetilde{\iota}_Y \iota^\dagger\right)\Delta - \widetilde{G}_Y G'\left(\widetilde{G}G' + (\epsilon n)^{-1} I\right)^{-1}\left((\epsilon n)^{-1} D (D'D/S)^{-1}\,\partial\Delta + \left(\widetilde{G} - \widetilde{\iota}\iota^\dagger\right)\Delta\right)\Bigg],$$

and (β,γ) are replaced by the preliminary $(\widehat{\beta},\widehat{\gamma})$ in all the quantities above, including when producing the simulated draws. $\widetilde{\delta}_\epsilon^{\text{MMSE}}$ is consistent for $\widehat{\delta}_\epsilon^{\text{MMSE}}$ as S tends to infinity for fixed n, under suitable regularity conditions (see Bonhomme, 2012, for a closely related setup). Note that matrix inverses remain well-defined as S tends to infinity, due to the presence of the Tikhonov-penalization term $(\epsilon n)^{-1} I$.

Confidence intervals. From Subsection 2.4, computing confidence intervals only requires, in addition to computing critical values under correct specification, to compute an estimate of the bias of the estimator $b_\epsilon(h,\widehat{\beta},\widehat{\gamma})$. In semi-parametric mixture models we have, for an asymptotically linear estimator based on h satisfying (2) and (4),

$$b_\epsilon(h,\beta_0,\gamma_*) = \epsilon^{\frac{1}{2}}\left\{\text{Var}_{\beta_0,\pi(\gamma_*)}[\Delta_{\beta_0}(A) - \mathbb{E}_{\beta_0,\pi(\gamma_*)}(h(Y)\,|\,A)]\right\}^{\frac{1}{2}}.$$

A numerical approximation to the bias of $\widehat{\delta}_\epsilon^{\text{MMSE}}$ is then

$$\widetilde{b}_\epsilon(h_\epsilon^{\text{MMSE}},\beta_0,\gamma_*) = \epsilon^{\frac{1}{2}}\left\|\Delta - \iota^\dagger\Delta - G'\widetilde{h}_\epsilon^{\text{MMSE}}\right\|.$$

Values of ϵ. In turn, ϵ_k in (29) can be approximated as $\mu(\alpha,p)^2/(n\lambda_k)$, where λ_k is the k-th largest eigenvalue of $G'QG = \widetilde{G}'\widetilde{G}$ (removing the eigenvalue equal to one since it corresponds to a constant eigenfunction).

S4 Models defined by moment restrictions

In this section, we consider settings where a finite-dimensional parameter $(\beta_0', \pi_0')'$ does not fully determine the distribution f_0 of Y, but satisfies a finite-dimensional system of moment conditions

$$\mathbb{E}_{f_0}\Psi(Y,\beta_0,\pi_0) = 0, \tag{S18}$$

which may be just-identified, over-identified or under-identified. We focus on asymptotically linear generalized method-of-moments (GMM) estimators of δ_{β_0,π_0} that satisfy

$$\widehat{\delta} = \delta_{\beta_0,\pi(\gamma_*)} + a(\beta_0,\gamma_*)'\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,\beta_0,\pi(\gamma_*)) + o_{P_0}(\epsilon^{\frac{1}{2}} + n^{-\frac{1}{2}}), \tag{S19}$$

for a parameter vector $a(\beta_0, \gamma_*)$. We will characterize the form of $a(\beta_0, \gamma_*)$ leading to minimum worst-case MSE in $\Gamma_\epsilon(\gamma_*)$.

We assume that the remainder in (S19) is uniformly bounded similarly as in (14). In this case local robustness with respect to $(\beta_0', \gamma_*')'$ takes the form

$$\nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{f_0}\nabla_{\beta\gamma}\Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*) = 0. \tag{S20}$$

It is natural to focus on asymptotically linear GMM estimators here, since f_0 is unrestricted except for the moment condition (S18).

To derive the worst-case bias of $\widehat{\delta}$ note that, by (S18), for any $\pi_0 \in \Gamma_\epsilon(\gamma_*)$ we have

$$\mathbb{E}_{f_0}\Psi(Y, \beta_0, \pi(\gamma_*)) = -\left[\mathbb{E}_{f_0}\nabla_\pi\Psi(Y, \beta_0, \pi(\gamma_*))\right]' (\pi_0 - \pi(\gamma_*)) + o(\epsilon^{\frac{1}{2}}),$$

so, under appropriate regularity conditions,

$$\sup_{\pi_0\in\Gamma_\epsilon(\gamma_*)} \left|\mathbb{E}_{f_0}\widehat{\delta} - \delta_{\beta_0,\pi_0}\right| = \epsilon^{\frac{1}{2}} \left\|\nabla_\pi\delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{f_0}\nabla_\pi\Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*)\right\|_{\gamma_*} + o(\epsilon^{\frac{1}{2}} + n^{-\frac{1}{2}}).$$

The worst-case MSE of

$$\widehat{\delta}_{a,\beta_0,\gamma_*} := \delta_{\beta_0,\pi(\gamma_*)} + a(\beta_0, \gamma_*)'\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i, \beta_0, \pi(\gamma_*))$$

is thus

$$\epsilon \left\|\nabla_\pi\delta_{\beta_0,\pi(\gamma_*)} + \mathbb{E}_{f_0}\nabla_\pi\Psi(Y, \beta_0, \pi(\gamma_*))\, a(\beta_0, \gamma_*)\right\|^2_{\gamma_*} + a(\beta_0, \gamma_*)'\frac{\mathbb{E}_{f_0}\Psi(Y, \beta_0, \pi(\gamma_*))\Psi(Y, \beta_0, \pi(\gamma_*))'}{n}a(\beta_0, \gamma_*) + o(\epsilon + n^{-1}).$$

To obtain an explicit expression for the minimum-MSE estimator, let us focus on the case where π_0 is finite-dimensional and $\|\cdot\|_{\gamma_*} = \|\cdot\|_{\Omega^{-1}}$. Let us define

$$V_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0}\Psi(Y, \beta_0, \pi(\gamma_*))\Psi(Y, \beta_0, \pi(\gamma_*))', \quad K_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0}\nabla_\pi\Psi(Y, \beta_0, \pi(\gamma_*)),$$

and

$$K_{\beta_0,\gamma_*} = \mathbb{E}_{f_0}\nabla_{\beta\gamma}\Psi(Y, \beta_0, \pi(\gamma_*)).$$

For all β_0, γ_* we aim to minimize

$$\epsilon \left\|\nabla_\pi\delta_{\beta_0,\pi(\gamma_*)} + K_{\beta_0,\pi(\gamma_*)}a(\beta_0, \gamma_*)\right\|^2_{\Omega^{-1}} + a(\beta_0, \gamma_*)'\frac{V_{\beta_0,\pi(\gamma_*)}}{n}a(\beta_0, \gamma_*),$$
$$\text{subject to } \nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)} + K_{\beta_0,\gamma_*}a(\beta_0, \gamma_*) = 0.$$

A solution is given by[6]

$$a_\epsilon^{\text{MMSE}}(\beta_0, \gamma_*) = -B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon}K'_{\beta_0,\gamma_*}\left(K_{\beta_0,\gamma_*}B^\dagger_{\beta_0,\pi(\gamma_*),\epsilon}K'_{\beta_0,\gamma_*}\right)^{-1}\nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)}$$

[6]Here we assume that $K_{\beta_0,\gamma_*}V^\dagger_{\beta_0,\pi(\gamma_*)}K'_{\beta_0,\gamma_*}$ is non-singular, requiring that β_0, γ_* be identified from the moment conditions. Existence follows from the fact that, by the generalized information identity, $V_{\beta_0,\pi(\gamma_*)}a = 0$ implies that $K_{\beta_0,\pi(\gamma_*)}a = 0$. Moreover, although $a_\epsilon^{\text{MMSE}}(\beta_0, \gamma_*)$ may not be unique, $a_\epsilon^{\text{MMSE}}(\beta_0, \gamma_*)'\Psi(Y, \beta_0, \pi(\gamma_*))$ is unique almost surely.

$$- B^{\dagger}_{\beta_0,\pi(\gamma_*),\epsilon} \left(I - K'_{\beta_0,\gamma_*} \left(K_{\beta_0,\gamma_*} B^{\dagger}_{\beta_0,\pi(\gamma_*),\epsilon} K'_{\beta_0,\gamma_*} \right)^{-1} K_{\beta_0,\gamma_*} B^{\dagger}_{\beta_0,\pi(\gamma_*),\epsilon} \right) K'_{\beta_0,\pi(\gamma_*)} \Omega^{-1} \nabla_{\pi} \delta_{\beta_0,\pi(\gamma_*)}, \tag{S21}$$

where $B_{\beta_0,\pi(\gamma_*),\epsilon} = K'_{\beta_0,\pi(\gamma_*)}\Omega^{-1}K_{\beta_0,\pi(\gamma_*)} + (\epsilon n)^{-1}V_{\beta_0,\pi(\gamma_*)}$, and $B^{\dagger}_{\beta_0,\pi(\gamma_*),\epsilon}$ is its Moore-Penrose generalized inverse. Note that, in the likelihood case and taking $\Psi(y,\beta,\pi) = \nabla_\pi \log f_{\beta,\pi}(y)$, the function $h(y,\beta_0,\gamma_*) = a^{\text{MMSE}}_{\epsilon}(\beta_0,\gamma_*)'\Psi(y,\beta_0,\pi(\gamma_*))$ simplifies to (20).

As a special case, when $\epsilon = 0$ we have

$$a^{\text{MMSE}}_0(\beta_0,\gamma_*) = -V^{\dagger}_{\beta_0,\pi(\gamma_*)}K'_{\beta_0,\gamma_*}\left(K_{\beta_0,\gamma_*}V^{\dagger}_{\beta_0,\pi(\gamma_*)}K'_{\beta_0,\gamma_*}\right)^{-1}\nabla_{\beta\gamma}\delta_{\beta_0,\pi(\gamma_*)}.$$

In this case, given preliminary estimators $\widehat{\beta}$ and $\widehat{\gamma}$, the minimum-MSE estimator

$$\widehat{\delta}^{\text{MMSE}}_{\epsilon} = \delta_{\widehat{\beta},\pi(\widehat{\gamma})} + a^{\text{MMSE}}_0(\widehat{\beta},\widehat{\gamma})'\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,\widehat{\beta},\pi(\widehat{\gamma}))$$

is the one-step approximation to the optimal GMM estimator based on the reference model. To obtain a feasible estimator one simply replaces the expectations in $V_{\beta_0,\pi(\gamma_*)}$ and K_{β_0,γ_*} by sample analogs.

As a second special case, consider ϵ tending to infinity. Focusing on the known-(β_0,γ_*) case for simplicity, $a^{\text{MMSE}}_{\epsilon}(\beta_0,\gamma_*)$ tends to $-K^{\text{ginv}}_{\beta_0,\pi(\gamma_*)}\nabla_\pi\delta_{\beta_0,\pi(\gamma_*)}$, where

$$K^{\text{ginv}}_{\beta_0,\pi(\gamma_*)} :=$$
$$\left(V^{\dagger}_{\beta_0,\pi(\gamma_*)}\right)^{1/2}\left[\left(V^{\dagger}_{\beta_0,\pi(\gamma_*)}\right)^{1/2}K'_{\beta_0,\pi(\gamma_*)}\Omega^{-1}K_{\beta_0,\pi(\gamma_*)}\left(V^{\dagger}_{\beta_0,\pi(\gamma_*)}\right)^{1/2}\right]^{\dagger}\left(V^{\dagger}_{\beta_0,\pi(\gamma_*)}\right)^{1/2}K'_{\beta_0,\pi(\gamma_*)}\Omega^{-1}$$

is a generalized inverse of $K_{\beta_0,\pi(\gamma_*)}$, and the choice of Ω corresponds to choosing one specific such generalized inverse. In this case, the minimum-MSE estimator is the one-step approximation to a particular GMM estimator based on the "large" model.

Lastly, given a parameter vector a, confidence intervals can be constructed as explained in Subsection 2.4, taking

$$b_{\epsilon}(a,\widehat{\beta},\widehat{\gamma}) = \epsilon^{\frac{1}{2}}\left\|\nabla_\pi\delta_{\widehat{\beta},\pi(\widehat{\gamma})} + \frac{1}{n}\sum_{i=1}^{n}\nabla_\pi\Psi(Y_i,\widehat{\beta},\pi(\widehat{\gamma}))\,a(\widehat{\beta},\widehat{\gamma})\right\|_{\Omega^{-1}}.$$

Example. Consider again the OLS/IV example of Subsection 3.3, but now drop the Gaussian assumptions on the distributions. For known C, the set of moment conditions corresponds to the moment functions

$$\Psi(y,x,z,\beta,\pi) = \begin{pmatrix} x(y - x'\beta - \pi'(x - Cz)) \\ z(y - x'\beta) \end{pmatrix}.$$

In this case, letting $W = (X',Z')'$ we have

$$K_{\beta_0,\gamma_*} = -\mathbb{E}_{f_0}(XW'), \quad K_{\beta_0,\pi(\gamma_*)} = -\mathbb{E}_{f_0}\begin{pmatrix} XX' & XZ' \\ (X - CZ)X' & 0 \end{pmatrix},$$

and

$$V_{\beta_0,\pi(\gamma_*)} = \mathbb{E}_{f_0}\left((Y - X'\beta_0)^2 WW'\right).$$

Given a preliminary estimator $\widetilde{\beta}$, $V_{\beta_0,\pi(\gamma_*)}$ can be estimated as $\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\widetilde{\beta})^2 W_i W_i'$, whereas K_{β_0,γ_*} and $K_{\beta_0,\pi(\gamma_*)}$ can be estimated as sample means. The estimator based on (S21) then interpolates nonlinearly between the OLS and IV estimators, similarly as in the likelihood case.

S5 Numerical illustrations

S5.1 Interpretation of ϵ in the cross-sectional binary choice model

Here we use the binary choice model of Subsection 6.1 to provide additional intuition about the interpretation of ϵ based on statistical testing.

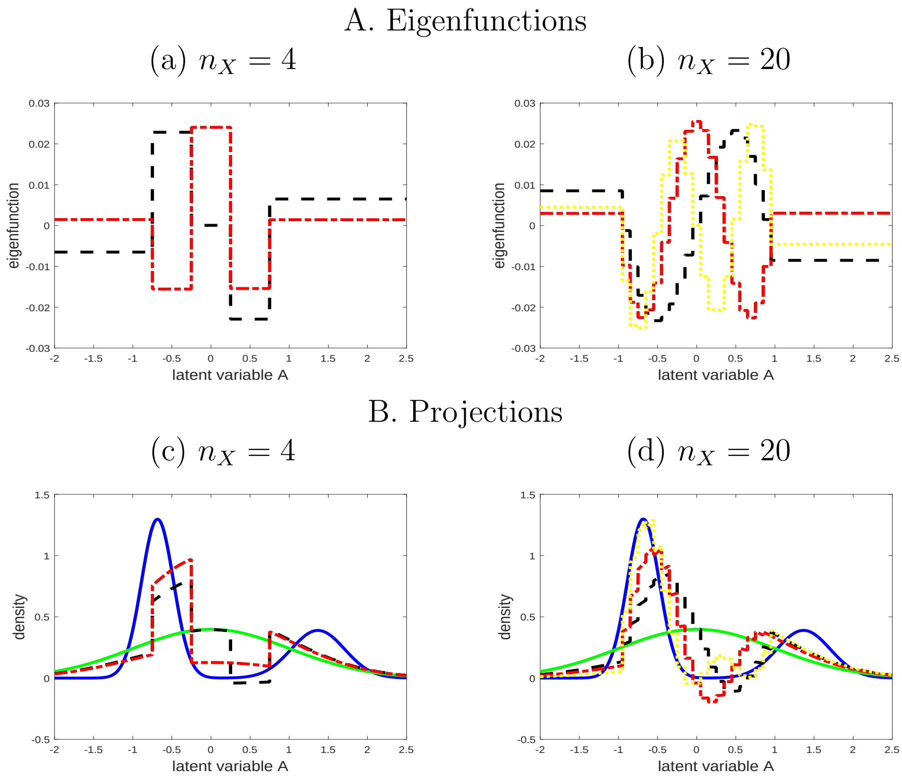
Let $\mathcal{U}_k$ denote the span of the first k non-constant eigenfunctions of the operator $\widetilde{H}_\pi$. By construction, any density $\pi_0 \notin \Gamma_{\epsilon_k}(\gamma_*)$ such that $(\pi_0 - \pi(\gamma_*))/\pi(\gamma_*) \in \mathcal{U}_k$ can be "detected" easily, in the sense that the local power of a 5%-likelihood ratio test exceeds 80%.[7] In the upper panel of Figure S1, we plot the eigenfunctions in $\mathcal{U}_k$. Plotting those allows one to visualize the directions along which setting ϵ to either of the ϵ_k's provides power guarantees outside the neighborhood. We see that the eigenfunctions do not vary outside the $[-1, 1]$ interval, where the support of $X'\beta_0$ lies. Within the $[-1, 1]$ interval, the eigenfunctions oscillate and belong to orthogonal bases of functions.

To see how well the true π_0 can be approximated using the directions in $\mathcal{U}_k$, in the bottom panel of Figure S1, we report the projection of π_0 onto $\mathcal{U}_k$. We see that, outside the $[-1, 1]$ interval, the projection is only governed by the reference normal density, reflecting the limited support of X. Within the interval, the approximation to the true bimodal density improves as k increases. At the same time, note that, consistently with our local approach, the approximating functions are not necessarily non-negative.[8]

[7] $\mathcal{U}_k$ consists of cotangent elements that have zero mean under the reference model. Any such $u \in \mathcal{T}$ can be mapped to a direction $v = u \cdot \pi(\gamma_*) \in \overline{\mathcal{T}}$ in the tangent space.

[8] In addition, since we know π_0 in this exercise, we can compute the local power of a 5%-likelihood ratio test in direction $\pi_0 - \pi(\gamma_*)$, for any value of ϵ. We find a power of 0.51 at ϵ_1 and 0.71 at ϵ_2 when X has 4 points of support, and 0.67 at ϵ_1, 0.92 at ϵ_2, and 0.99 at ϵ_3 when X has 20 points of support.

Figure S1: Eigenfunctions of $\widetilde{H}_\pi$ in the cross-sectional binary choice model



Notes: In the top panel we report the first 2 (respectively, first 3) non-constant eigenfunctions of $\widetilde{H}_\pi$*. The first eigenfunction is shown in dashed, the second one in dashed-dotted, and the third one in dotted. In the bottom panel we plot the true and reference densities in solid, as well as the successive approximations using the first, the first two, or the first three eigenfunctions.*

S5.2 Additional tables

Table S1: Monte Carlo simulation of the average effect in the cross-sectional binary choice model, *interpolation* ($x_0 = (0.5, 1)'$)

Minimum-MSE, for $\epsilon =$	0.0001	0.20	0.40	0.60	0.80	1.00
			A. $n_X = 4$			
Worst-case bias	0.0021	0.0783	0.1104	0.1351	0.1560	0.1744
Asymptotic standard error	0.0228	0.0288	0.0297	0.0300	0.0302	0.0303
Monte Carlo bias	0.1026	0.0197	0.0134	0.0111	0.0099	0.0092
Monte Carlo standard deviation	0.0253	0.0281	0.0288	0.0291	0.0292	0.0293
Monte Carlo root MSE	0.1057	0.0343	0.0317	0.0311	0.0308	0.0307
CI length	0.0936	0.2697	0.3372	0.3878	0.4302	0.4674
CI coverage	0.0180	0.9990	1.0000	1.0000	1.0000	1.0000
			B. $n_X = 20$			
Worst-case bias	0.0021	0.0480	0.0610	0.0714	0.0805	0.0887
Asymptotic standard error	0.0227	0.0394	0.0453	0.0487	0.0509	0.0526
Monte Carlo bias	0.0976	0.0080	0.0037	0.0026	0.0022	0.0020
Monte Carlo standard deviation	0.0239	0.0386	0.0446	0.0480	0.0502	0.0519
Monte Carlo root MSE	0.1005	0.0394	0.0447	0.0480	0.0502	0.0519
CI length	0.0931	0.2503	0.2996	0.3337	0.3607	0.3835
CI coverage	0.0190	0.9990	1.0000	1.0000	1.0000	1.0000

Notes: Performance of the minimum-MSE estimator in the cross-sectional binary choice model, for different values of ϵ*.* $n = 500$*, results for* 1000 *simulations. The nominal level for confidence intervals (CI) is 95%.* n_X *denotes the number of points of support of the first component of* X*.*

Table S2: Monte Carlo simulation of the average effect in the cross-sectional binary choice model, *extrapolation* ($x_0 = (-0.5, 1)'$)

Minimum-MSE, for $\epsilon =$	0.0001	0.20	0.40	0.60	0.80	1.00
			A. $n_X = 4$			
Worst-case bias	0.0029	0.1269	0.1794	0.2197	0.2537	0.2837
Asymptotic standard error	0.0296	0.0312	0.0315	0.0316	0.0316	0.0317
Monte Carlo bias	-0.0987	-0.0903	-0.0901	-0.0900	-0.0900	-0.0900
Monte Carlo standard deviation	0.0283	0.0330	0.0334	0.0335	0.0336	0.0336
Monte Carlo root MSE	0.1027	0.0961	0.0961	0.0961	0.0961	0.0961
CI length	0.1219	0.3762	0.4822	0.5632	0.6314	0.6914
CI coverage	0.2000	0.9370	0.9850	0.9960	0.9990	1.0000
			B. $n_X = 20$			
Worst-case bias	0.0028	0.1172	0.1645	0.2008	0.2314	0.2584
Asymptotic standard error	0.0313	0.0401	0.0443	0.0470	0.0489	0.0503
Monte Carlo bias	-0.0902	-0.0961	-0.0988	-0.0999	-0.1005	-0.1009
Monte Carlo standard deviation	0.0287	0.0373	0.0412	0.0437	0.0456	0.0471
Monte Carlo root MSE	0.0947	0.1031	0.1070	0.1090	0.1104	0.1113
CI length	0.1284	0.3915	0.5026	0.5857	0.6544	0.7141
CI coverage	0.2530	0.9500	0.9910	0.9960	0.9970	0.9970

Notes: See the notes to Table S1.

Table S3: Monte Carlo simulation results for the autoregressive parameter in the dynamic binary choice panel data model

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
	A. $T = 5$					
Worst-case bias	0.0001	0.0179	0.0227	0.0266	0.0299	0.0327
Asymptotic standard error	0.0952	0.0975	0.0979	0.0981	0.0983	0.0985
Monte Carlo bias	-0.1729	-0.0615	-0.0555	-0.0531	-0.0518	-0.0509
Monte Carlo standard deviation	0.1252	0.1111	0.1129	0.1136	0.1141	0.1145
Monte Carlo root MSE	0.2135	0.1270	0.1258	0.1255	0.1254	0.1253
CI length	0.3734	0.4179	0.4292	0.4379	0.4452	0.4516
CI coverage	0.5470	0.8890	0.9080	0.9160	0.9220	0.9280
	B. $T = 10$					
Worst-case bias	0.0001	0.0090	0.0118	0.0140	0.0158	0.0175
Asymptotic standard error	0.0607	0.0614	0.0615	0.0616	0.0616	0.0617
Monte Carlo bias	-0.0780	-0.0137	-0.0120	-0.0114	-0.0110	-0.0107
Monte Carlo standard deviation	0.0676	0.0731	0.0736	0.0738	0.0739	0.0740
Monte Carlo root MSE	0.1032	0.0744	0.0745	0.0746	0.0747	0.0748
CI length	0.2381	0.2587	0.2647	0.2694	0.2733	0.2768
CI coverage	0.7130	0.9210	0.9330	0.9360	0.9360	0.9380
	C. $T = 20$					
Worst-case bias	0.0001	0.0058	0.0078	0.0093	0.0106	0.0118
Asymptotic standard error	0.0418	0.0421	0.0422	0.0422	0.0422	0.0422
Monte Carlo bias	-0.0304	-0.0023	-0.0019	-0.0017	-0.0017	-0.0016
Monte Carlo standard deviation	0.0442	0.0488	0.0490	0.0490	0.0491	0.0491
Monte Carlo root MSE	0.0537	0.0488	0.0490	0.0491	0.0491	0.0491
CI length	0.1638	0.1766	0.1808	0.1840	0.1867	0.1891
CI coverage	0.8780	0.9110	0.9180	0.9230	0.9260	0.9300

Notes: Performance of the minimum-MSE estimator of β_0 *in the dynamic panel data binary choice model, for different values of* ϵ*.* $n = 500$*, results for* 1000 *simulations. The nominal level for confidence intervals (CI) is 95%.*

Table S4: Monte Carlo simulation results for the average state dependence parameter in the dynamic binary choice panel data model

Minimum-MSE, for $\epsilon =$	0.00	0.20	0.40	0.60	0.80	1.00
			A. $T = 5$			
Worst-case bias	0.0000	0.0099	0.0134	0.0162	0.0185	0.0205
Asymptotic standard error	0.0259	0.0268	0.0270	0.0272	0.0273	0.0274
Monte Carlo bias	-0.0538	-0.0218	-0.0202	-0.0196	-0.0193	-0.0191
Monte Carlo standard deviation	0.0439	0.0324	0.0331	0.0334	0.0336	0.0337
Monte Carlo root MSE	0.0694	0.0391	0.0387	0.0387	0.0387	0.0388
CI length	0.1017	0.1250	0.1329	0.1389	0.1439	0.1483
CI coverage	0.4450	0.8620	0.8850	0.9000	0.9190	0.9240
			B. $T = 10$			
Worst-case bias	0.0000	0.0121	0.0169	0.0207	0.0238	0.0266
Asymptotic standard error	0.0181	0.0184	0.0185	0.0186	0.0186	0.0187
Monte Carlo bias	-0.0212	-0.0047	-0.0048	-0.0050	-0.0051	-0.0052
Monte Carlo standard deviation	0.0257	0.0229	0.0230	0.0231	0.0231	0.0232
Monte Carlo root MSE	0.0333	0.0233	0.0235	0.0236	0.0237	0.0238
CI length	0.0710	0.0963	0.1063	0.1141	0.1206	0.1263
CI coverage	0.6610	0.9630	0.9780	0.9830	0.9870	0.9880
			C. $T = 20$			
Worst-case bias	0.0000	0.0163	0.0230	0.0281	0.0325	0.0363
Asymptotic standard error	0.0134	0.0135	0.0136	0.0137	0.0137	0.0138
Monte Carlo bias	-0.0097	-0.0028	-0.0028	-0.0028	-0.0028	-0.0028
Monte Carlo standard deviation	0.0187	0.0153	0.0153	0.0154	0.0154	0.0155
Monte Carlo root MSE	0.0210	0.0155	0.0156	0.0156	0.0157	0.0157
CI length	0.0525	0.0857	0.0993	0.1098	0.1187	0.1265
CI coverage	0.7840	0.9890	0.9930	0.9960	0.9970	0.9970

Notes: Performance of the minimum-MSE estimator of δ_{β_0,π_0} in the dynamic panel data binary choice model, for different values of ϵ. $n = 500$, results for 1000 simulations. The nominal level for confidence intervals (CI) is 95%.

References

[1] Bonhomme, S. (2012): "Functional Differencing," *Econometrica*, 80(4), 1337–1385.

[2] Dharmadhikari, S. W., and Jogdeo, K. (1969): "Bounds on Moments of Certain Random Variables," *The Annals of Mathematical Statistics*, 40(4), 1506-1509.

[3] Engl, H.W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems*, Kluwer Academic Publishers.

[4] Kress, R. (2014): *Linear Integral Equations*. Applied Mathematical Sciences. Volume 82. Springer. New York.

[5] Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer Verlag Berlin, Heidelberg, New York.

[6] Van der Vaart, A. W. (2007): *Asymptotic Statistics*, Cambridge University Press.