MODEL AVERAGING BASED ON KULLBACK-LEIBLER DISTANCE

Xinyu Zhang[1], Guohua Zou[1,2] and Raymond J. Carroll[3]

[1]*Chinese Academy of Sciences,* [2]*Capital Normal University and* [3]*Texas A&M University*

Supplementary Material

This file contains proofs and additional simulation results as follows:

- Section S1: Proof of Theorem 1;
- Section S2: Proof of Theorem 2;
- Section S3: Proof of Theorem 3;
- Section S4: Proof of (2.4);
- Section S5: Proof of Theorem 4;
- Section S6: Calculation for $\partial\widehat{\eta}^{\mathrm{T}}/\partial y$;
- Section S7: Discussion on Assumption (A.1) and its relationship with the normality of e;
- Section S8: All Simulation Results of Example 1;
- Section S9: All Simulation Results of Example 2;
- Section S10: All Simulation Results of Example 3;
- Section S11: Simulation Results with a Uniformly Distributed Error Term;
- Section S12: Simulation Results with a Chi-squared Distributed Error Term;
- Section S13: Simulation Results with Coefficients Depending on the Sample Size.

S1 Proof of Theorem 1

This proof is mainly an application of Stein's Lemma (Stein (1981)). Let y_i, μ_i and $\widehat{\mu}(w)_i$ be the i^{th} elements of y, μ and $\widehat{\mu}(w)$, respectively. By the conditions stated in Theorem 1 and Stein's Lemma, we have

$$E_{f(y)}\left[\frac{(y_i-\mu_i)\left\{y_i-\widehat{\mu}(w)_i\right\}}{\widehat{\sigma}^2}\right]=\sigma^2 E_{f(y)}\left(\frac{\partial\left[\left\{y_i-\widehat{\mu}(w)_i\right\}/\widehat{\sigma}^2\right]}{\partial y_i}\right) \tag{S1.1}$$

and

$$E_{f(y)}\left\{\frac{(y_i-\mu_i)^2}{\widehat{\sigma}^2}\right\}=\sigma^2E_{f(y)}(\widehat{\sigma}^{-2})+\sigma^4E_{f(y)}\left\{\frac{\partial^2(\widehat{\sigma}^{-2})}{\partial y_i^2}\right\}. \tag{S1.2}$$

Using (S1.1), it can be seen that

$$\begin{aligned}E_{f(y)}\left[\frac{(y-\mu)^{\mathrm{T}}\{y-\widehat{\mu}(w)\}}{\widehat{\sigma}^2}\right] &= \sigma^2E_{f(y)}\left(\text{trace}\frac{\partial\left[\{y-\widehat{\mu}(w)\}/\widehat{\sigma}^2\right]}{\partial y^{\mathrm{T}}}\right)\\ &= E_{f(y)}(n\sigma^2\widehat{\sigma}^{-2})-E_{f(y)}\text{trace}\left\{\sigma^2\widehat{\sigma}^{-2}\frac{\partial\widehat{\mu}(w)}{\partial y^{\mathrm{T}}}\right\}\\ &\quad -E_{f(y)}\left[\sigma^2\widehat{\sigma}^{-4}\{y-\widehat{\mu}(w)\}^{\mathrm{T}}\frac{\partial\widehat{\sigma}^2}{\partial y}\right].\end{aligned} \tag{S1.3}$$

Using (S1.2), we obtain

$$\begin{aligned}&E_{f(y)}\left(\frac{\|y-\mu\|^2}{\widehat{\sigma}^2}\right)=n\sigma^2E_{f(y)}\left(\widehat{\sigma}^{-2}\right)+\sigma^4E_{f(y)}\left[\text{trace}\left\{\frac{\partial^2(\widehat{\sigma}^{-2})}{\partial y\partial y^{\mathrm{T}}}\right\}\right]\\ &=E_{f(y)}\left(n\sigma^2\widehat{\sigma}^{-2}\right)+E_{f(y)}\left\{2\sigma^4\widehat{\sigma}^{-6}\text{trace}\left(\frac{\partial\widehat{\sigma}^2}{\partial y}\frac{\partial\widehat{\sigma}^2}{\partial y^{\mathrm{T}}}\right)\right\}\\ &\quad -E_{f(y)}\left\{\sigma^4\widehat{\sigma}^{-4}\text{trace}\left(\frac{\partial^2\widehat{\sigma}^2}{\partial y\partial y^{\mathrm{T}}}\right)\right\}.\end{aligned} \tag{S1.4}$$

In addition, it is straightforward to show that

$$\|\mu-\widehat{\mu}(w)\|^2=\|y-\widehat{\mu}(w)\|^2+\|y-\mu\|^2-2\{y-\widehat{\mu}(w)\}^{\mathrm{T}}(y-\mu). \tag{S1.5}$$

Now, by combining (S1.3)-(S1.5), Theorem 1 is proved.

S2 Proof of Theorem 2

Based on the proofs of Theorems 1' and 2 in Wan, Zhang and Zou (2010), and Assumptions (A.1), (A.2) and (A.4), to prove Theorem 2, we need only to verify that

$$\sup_{w\in\mathcal{W}}\left\{\left|y^{\mathrm{T}}P^{\mathrm{T}}(w)\frac{\partial\widehat{\sigma}^2}{\partial y}\right|R_n^{-1}(w)\right\}=o_p(1). \tag{S2.1}$$

By Assumptions (A.2), (A.3), and (A.5), we have

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left\{\left|y^{\mathrm{T}}P^{\mathrm{T}}(w)\frac{\partial\widehat{\sigma}^2}{\partial y}\right|R_n^{-1}(w)\right\} \le \xi_n^{-1}\sup_{w\in\mathcal{W}}\left|y^{\mathrm{T}}P^{\mathrm{T}}(w)\widehat{T}y\right| \\
&\quad = \xi_n^{-1}\sup_{w\in\mathcal{W}}\left|y^{\mathrm{T}}\left\{P^{\mathrm{T}}(w)\widehat{T}+\widehat{T}^{\mathrm{T}}P(w)\right\}y\right|/2 \\
&\quad \le \xi_n^{-1}\sup_{w\in\mathcal{W}}\left[\lambda_{\max}\left\{P^{\mathrm{T}}(w)\widehat{T}+\widehat{T}^{\mathrm{T}}P(w)\right\}\right]\|y\|^2/2 \\
&\quad \le \xi_n^{-1}\sup_{w\in\mathcal{W}}\left[\lambda_{\max}\{P(w)\}\right]\lambda_{\max}(\widehat{T})\|y\|^2 \\
&\quad \le \xi_n^{-1}\sup_{w\in\mathcal{W}}\left\{\max_{s\in\{1,\ldots,S\}}\lambda_{\max}(P_{(s)})\right\}\lambda_{\max}(\widehat{T})\|y\|^2 \\
&\quad = \xi_n^{-1}\max_{s\in\{1,\ldots,S\}}\lambda_{\max}(P_{(s)})\cdot n\lambda_{\max}(\widehat{T})\cdot n^{-1}\|y\|^2 = o_p(1). \qquad \text{(S2.2)}
\end{aligned}$$

This completes the proof.

S3 Proof of Theorem 3

By the assumption that μ is a linear function of X, we have

$$y^{\mathrm{T}}(I_n-P)y = e^{\mathrm{T}}(I_n-P)e,$$

where $I_n - P$ is a symmetric idempotent matrix with rank $n-m$. So $\sigma^2/\{y^{\mathrm{T}}(I_n-P)y\}$ is distributed as an inverse Chi-squared distribution with mean $(n-m-2)^{-1}$, and thus

$$E_{f(y)}\left\{\frac{\sigma^2}{\widehat{\sigma}^2(y,k)}\right\} = \frac{k}{n-m-2}. \qquad \text{(S3.1)}$$

In addition,

$$\begin{aligned}
\{y-\widehat{\mu}(w)\}^{\mathrm{T}}\frac{\partial\widehat{\sigma}^2(y,k)}{\partial y} &= 2k^{-1}\{y-\widehat{\mu}(w)\}^{\mathrm{T}}(I_n-P)y \\
&= 2k^{-1}y^{\mathrm{T}}\sum_{s=1}^{S}w_s(I_n-P_{(s)})(I_n-P)y \\
&= 2k^{-1}y^{\mathrm{T}}(I_n-P)y. \qquad \text{(S3.2)}
\end{aligned}$$

By applying (S3.1), (S3.2), and the facts that $\partial\widehat{\sigma}^2(y,k)/\partial y = 2k^{-1}(I_n-P)y$ and $\partial^2\widehat{\sigma}^2(y,k)/(\partial y\partial y^{\mathrm{T}})$ $= 2k^{-1}(I_n-P)$ to Theorem 1, we obtain Theorem 3.

S4 Proof of (2.4)

Let $\widetilde{\mathcal{C}}^*(w) = \mathcal{C}^*(w) - \|e\|^2$ such that $\widehat{w} = \underset{w\in\mathcal{W}}{\text{argmin}}\widetilde{\mathcal{C}}^*(w)$. Write $\widetilde{w} = \underset{w\in\mathcal{W}}{\text{argmin}}E\{L_n(w)\}$. It is straightforward to show that

$$\begin{aligned} L_n(\widehat{w}) &= \widetilde{\mathcal{C}}^*(\widehat{w}) - c(\widehat{w}) \\ &= \inf_{w\in\mathcal{W}} \widetilde{\mathcal{C}}^*(w) - c(\widehat{w}) \\ &= \inf_{w\in\mathcal{W}} \{L_n(w) + c(w)\} - c(\widehat{w}) \\ &\leq L_n(\widetilde{w}) + e'\{I_n - P(\widetilde{w})\}\mu + \sigma^2 \text{trace}\{P(\widetilde{w})\} - e'P(\widetilde{w})e - c(\widehat{w}), \end{aligned}$$

which implies (2.4).

S5 Proof of Theorem 4

Let $\widetilde{y} = \Omega^{-1/2}y$, $\widetilde{\mu} = \Omega^{-1/2}\mu$, $\widetilde{e} = \Omega^{-1/2}e$, and $\widetilde{\Omega} = \Omega^{-1/2}\widehat{\Omega}\Omega^{-1/2}$. By simple calculations, we have

$$\begin{aligned} \mathcal{D}^*(w) &= \{\widetilde{y} - \widetilde{P}(w)\widetilde{y})\}^{\mathrm{T}}\{\widetilde{y} - \widetilde{P}(w)\widetilde{y})\} + 2\text{trace}\{\widetilde{P}(w)\} \\ &+\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\} + \{\widetilde{P}(w)\widetilde{e}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e} \\ &-2\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e} + 2\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\} \\ &-2\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e} - 2\widetilde{y}^{\mathrm{T}}\widetilde{P}^{\mathrm{T}}(w)\widetilde{\Omega}^{-1}\Omega^{-1/2}\widehat{a} + \widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{e}. \end{aligned} \tag{S5.1}$$

It follows from Assumptions (A.2) and (B.2) that

$$\sup_{w\in\mathcal{W}} \lambda_{\max}\{\widetilde{P}(w)\} = O(1). \tag{S5.2}$$

From the proof of Theorem 1' in Wan, Zhang and Zou (2010), (S5.1), (S5.2), Assumptions (B.1)-(B.2), and the fact that $\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{e}$ is unrelated to w, in order to prove (3.2), we need only to verify that

$$\sup_{w\in\mathcal{W}} [R^{-1}_{\text{hetero},n}(w)|\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}|] = o_p(1), \tag{S5.3a}$$

$$\sup_{w\in\mathcal{W}} [R^{-1}_{\text{hetero},n}(w)|\{\widetilde{P}(w)\widetilde{e}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}|] = o_p(1), \tag{S5.3b}$$

$$\sup_{w\in\mathcal{W}} [R^{-1}_{\text{hetero},n}(w)|\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}|] = o_p(1), \tag{S5.3c}$$

$$\sup_{w\in\mathcal{W}} [R^{-1}_{\text{hetero},n}(w)|\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}|] = o_p(1), \tag{S5.3d}$$

$$\sup_{w\in\mathcal{W}} \{R^{-1}_{\text{hetero},n}(w)|\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}|\} = o_p(1), \tag{S5.3e}$$

$$\sup_{w\in\mathcal{W}} \{R^{-1}_{\text{hetero},n}(w)|\widetilde{y}^{\mathrm{T}}\widetilde{P}^{\mathrm{T}}(w)\widetilde{\Omega}^{-1}\Omega^{1/2}\widehat{a}|\} = o_p(1). \tag{S5.3f}$$

By (S5.2) and Assumptions (A.3) and (B.2), it is straightforward to show that

$$\begin{aligned} R_{\text{hetero},n}(w) &= \|\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\|^2 + \text{trace}\{\widetilde{P}(w)\widetilde{P}^{\mathrm{T}}(w)\} \\ &\geq \max\{\|\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\|^2, \text{trace}\{\widetilde{P}(w)\widetilde{P}^{\mathrm{T}}(w)\}\}, \qquad \text{(S5.4)} \\ R_{\text{hetero},n}(w) &= \mu^{\mathrm{T}}\{I_n - P(w)\}^{\mathrm{T}}\Omega^{-1}\{I_n - P(w)\}\mu + \text{trace}\{P(w)\Omega P^{\mathrm{T}}(w)\Omega^{-1}\} \\ &= O(n), \qquad \text{(S5.5)} \end{aligned}$$

and that

$$\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\|\widetilde{P}(w)\widetilde{e}\|^2\} = o_p(1) + \sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)\text{trace}\{\widetilde{P}(w)\widetilde{P}^{\mathrm{T}}(w)\}] \qquad \text{(S5.6)}$$

by the proof of Theorem 1' in Wan, Zhang and Zou (2010).

Using (S5.5), we see that $n^{-1}\xi_{\text{hetero},n} = O(1)$, which together with Assumption (B.3) implies that

$$\max_{i\in\{1,\ldots,n\}}|\widehat{\Omega}_{ii} - \Omega_{ii}| = o_p(1). \qquad \text{(S5.7)}$$

Using (S5.7) and Assumption (B.2), we have

$$\lambda_{\max}(\widetilde{\Omega}^{-1} - I_n) = O_p(\max_{i\in\{1,\ldots,n\}}|\widehat{\Omega}_{ii} - \Omega_{ii}|) = o_p(1). \qquad \text{(S5.8)}$$

Thus, from (S5.4) and (S5.8), we obtain

$$\begin{aligned} &\sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)|\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu})\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}|] \\ &\quad\leq \lambda_{\max}(\widetilde{\Omega}^{-1} - I_n)\sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu})\}^{\mathrm{T}}\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}] \\ &\quad\leq \lambda_{\max}(\widetilde{\Omega}^{-1} - I_n) = o_p(1), \end{aligned}$$

which is the result (S5.3a). Using (S5.6), (S5.8), and Assumption (B.4), it is seen that

$$\begin{aligned} &\sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)|\{\widetilde{P}(w)\widetilde{e}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}|] \\ &\quad\leq \lambda_{\max}(\widetilde{\Omega}^{-1} - I_n)\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\|\widetilde{P}(w)\widetilde{e}\|^2\} \\ &\quad= \lambda_{\max}(\widetilde{\Omega}^{-1} - I_n)\left(\sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)\text{trace}\{\widetilde{P}(w)\widetilde{P}^{\mathrm{T}}(w)\}] + o_p(1)\right) \\ &\quad= o_p(1), \qquad \text{(S5.9)} \end{aligned}$$

which is the result (S5.3b). Using (S5.4), (S5.6), (S5.8), and Assumption (B.4), we obtain

$$\begin{aligned} &\sup_{w\in\mathcal{W}}(R^{-2}_{\text{hetero},n}(w)[\{\widetilde{\mu} - \widetilde{P}(w)\widetilde{\mu}\}^{\mathrm{T}}(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}]^2) \\ &\quad\leq \sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\|(\widetilde{\Omega}^{-1} - I_n)\widetilde{P}(w)\widetilde{e}\|^2\} \\ &\quad\leq \{\lambda_{\max}(\widetilde{\Omega}^{-1} - I_n)\}^2\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\|\widetilde{P}(w)\widetilde{e}\|^2\} = o_p(1), \end{aligned}$$

which implies the result (S5.3c). Using (S5.4), (S5.8), and Assumptions (B.2)-(B.3), it follows that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}[R^{-2}_{\text{hetero},n}(w)\{\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1}-I_n)\{\widetilde{\mu}-\widetilde{P}(w)\widetilde{\mu})\}^2]\\
&\quad\le\{\lambda_{\max}(\widetilde{\Omega}^{-1}-I_n)\}^2\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\}\|\widetilde{e}\|^2\\
&\quad\le\{\lambda_{\max}(\widetilde{\Omega}^{-1}-I_n)\}^2 n\xi^{-1}_{\text{hetero},n}n^{-1}\|\widetilde{e}\|^2=o_p(1),
\end{aligned}$$

which implies the result (S5.3d). Similarly, from (S5.4), (S5.6), (S5.8), and Assumptions (B.2)-(B.3), we have

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}[R^{-2}_{\text{hetero},n}(w)\{\widetilde{e}^{\mathrm{T}}(\widetilde{\Omega}^{-1}-I_n)\widetilde{P}(w)\widetilde{e}\}^2]\\
&\quad\le\{\lambda_{\max}(\widetilde{\Omega}^{-1}-I_n)\}^2\xi^{-1}_{\text{hetero},n}\|\widetilde{e}\|^2\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)\|\widetilde{P}(w)\widetilde{e}\|^2\}\\
&\quad=o_p(1),
\end{aligned}$$

which implies the result (S5.3e). Finally, using steps similar to (S2.2), it follows from (S5.2), (S5.7), and Assumptions (A.3), (B.2) and (B.5) that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)|\widetilde{y}^{\mathrm{T}}\widetilde{P}(w)\widetilde{\Omega}^{-1}\Omega^{1/2}\widehat{a}|\}=\sup_{w\in\mathcal{W}}\{R^{-1}_{\text{hetero},n}(w)|\widetilde{y}^{\mathrm{T}}\widetilde{P}(w)\widetilde{\Omega}^{-1}\Omega^{1/2}\widehat{A}\widetilde{y}|\}\\
&\quad=\sup_{w\in\mathcal{W}}[R^{-1}_{\text{hetero},n}(w)|\widetilde{y}^{\mathrm{T}}\{\widetilde{P}(w)\widetilde{\Omega}^{-1}\Omega^{1/2}\widehat{A}+\widehat{A}^{\mathrm{T}}\Omega^{1/2}\widetilde{\Omega}^{-1}\widetilde{P}(w)^{\mathrm{T}}\}\widetilde{y}|]/2\\
&\quad\le\lambda_{\max}\{\widetilde{P}(w)\}\lambda_{\max}(\widetilde{\Omega}^{-1})\lambda_{\max}(\Omega^{1/2})n^{-1}\|\widetilde{y}\|^2 n\lambda_{\max}(\widehat{A})\xi^{-1}_{\text{hetero},n}\\
&\quad=o_p(1),
\end{aligned}$$

which is the result (S5.3f). This completes the proof.

S6 Calculation for $\partial\widehat{\eta}^{\mathrm{T}}/\partial y$

Assuming that $\mu=X\beta$, the -2log-likelihood is

$$n\log 2\pi+\log|\Omega(\eta)|+(y-X\beta)^{\mathrm{T}}\Omega^{-1}(\eta)(y-X\beta).$$

Write $\widehat{\eta}=(\widehat{\eta}_1,\ldots,\widehat{\eta}_q)^{\mathrm{T}}$. Let $\widehat{\beta}$ be the ML estimator of β, $U_j(\widehat{\eta})=\partial\Omega^{-1}(\widehat{\eta})/\partial\widehat{\eta}_j$, and $V_{jk}(\widehat{\eta})=\partial U_j(\widehat{\eta})/\partial\widehat{\eta}_k$ for $j,k=1,\ldots,q$. From Magnus and Neudecker (1988), we have the following score equations:

$$\begin{cases}
(y-X\widehat{\beta})^{\mathrm{T}}\Omega^{-1}(\widehat{\eta})X=0,\\
(y-X\widehat{\beta})^{\mathrm{T}}U_j(\widehat{\eta})(y-X\widehat{\beta})-\text{trace}\{U_j(\widehat{\eta})\Omega(\widehat{\eta})\}=0, \qquad j=1,...,q.
\end{cases} \tag{S6.1}$$

Let K be a $q\times n$ matrix with the j^{th} row $K_j=(y-X\widehat{\beta})^{\mathrm{T}}U_j(\widehat{\eta})$, Q be a $q\times q$ matrix with the jk^{th} element $Q_{jk}=(y-X\widehat{\beta})^{\mathrm{T}}V_{jk}(\widehat{\eta})(y-X\widehat{\beta})$, and B be a $q\times q$ matrix with the jk^{th} element $B_{jk}=\partial\text{trace}\{U_j(\widehat{\eta})\Omega(\widehat{\eta})\}/\partial\widehat{\eta}_k$.

Taking derivatives with respect to y on both sides of (S6.1), we have

$$\left\{\begin{array}{l} \frac{\partial\widehat{\eta}^{\mathrm{T}}}{\partial y}KX+\Omega^{-1}(\widehat{\eta})X-\frac{\partial\widehat{\beta}^{\mathrm{T}}}{\partial y}X^{\mathrm{T}}\Omega^{-1}(\widehat{\eta})X=0, \\ \frac{\partial\widehat{\eta}^{\mathrm{T}}}{\partial y}Q^{\mathrm{T}}+2K^{\mathrm{T}}-2\frac{\partial\widehat{\beta}^{\mathrm{T}}}{\partial y}X^{\mathrm{T}}K^{\mathrm{T}}-\frac{\partial\widehat{\eta}^{\mathrm{T}}}{\partial y}B^{\mathrm{T}}=0, \end{array}\right. \tag{S6.2}$$

by which, we have

$$\begin{aligned} \frac{\partial\widehat{\eta}^{\mathrm{T}}}{\partial y} &= [2\Omega^{-1}(\widehat{\eta})X\{X^{\mathrm{T}}\Omega^{-1}(\widehat{\eta})X\}^{-1}X^{\mathrm{T}}K^{\mathrm{T}}-2K^{\mathrm{T}}] \\ &\quad\times[Q^{\mathrm{T}}-B^{\mathrm{T}}-2KX\{X^{\mathrm{T}}\Omega^{-1}(\widehat{\eta})X\}^{-1}X^{\mathrm{T}}K^{\mathrm{T}}]^{-1}, \end{aligned}$$

given the above inverses exist.

S7 Discussion on Assumption (A.1) and its relationship with the normality of e

It is seen that

$$\begin{aligned} R_n(w) &= E\{L_n(w)\}=E\{\|\widehat{\mu}(w)-\mu\|^2\}=E\{\|P(w)(\mu+e)-\mu\|^2\} \\ &= \|P(w)\mu-\mu\|^2+\sigma^2\mathrm{trace}\{P(w)P^{\mathrm{T}}(w)\}, \end{aligned}$$

which depends on the mean and variance of e_i, but is unrelated to other properties of e_i. So given the mean zero and the variance σ^2, Assumption (A.1) depends on the distribution of e_i only through the integer G, which is determined by the first part of Assumption (A.1). The first part of Assumption (A.1) is just a moment condition, which is satisfied for any $G<\infty$ when e_i follows normal, uniform or Chi-squared distribution. Thus, if Assumption (A.1) holds for normal e_i, then it also holds for uniform or Chi-squared e_i.

In the following, we discuss the assumption

$$S\xi_n^{-2G}\sum_{s=1}^{S}R_n^G(w_s^0)=o(1). \tag{S7.1}$$

First, $\xi_n\to\infty$ is a necessary condition of the assumption (S7.1). As Hansen and Racine (2012) remarked, this condition requires that all finite dimensional models be approximations. Let $\eta_n=\max_{s\in\{1,\ldots,S\}}R_n(w_s^0)$. A sufficient condition of the assumption (S7.1) is $S^2(\xi_n^{-2}\eta_n)^G=o(1)$. For a fixed rate of $\xi_n\to\infty$, the slower the rates of $S\to\infty$ and $\eta_n\to\infty$, the faster the rate of $S^2(\xi_n^{-2}\eta_n)^G\to 0$. Practically, the rates of $S\to\infty$ and $\eta_n\to\infty$ can be reduced by removing the very poor models at the outset prior to model averaging.

In fact, the assumption (S7.1) is introduced by Wan, Zhang and Zou (2010), where its rationality was discussed in detail. Since that, this condition has been used in model averaging studies such as Liu and Okui (2013).

S8 All Simulation Results of Example 1

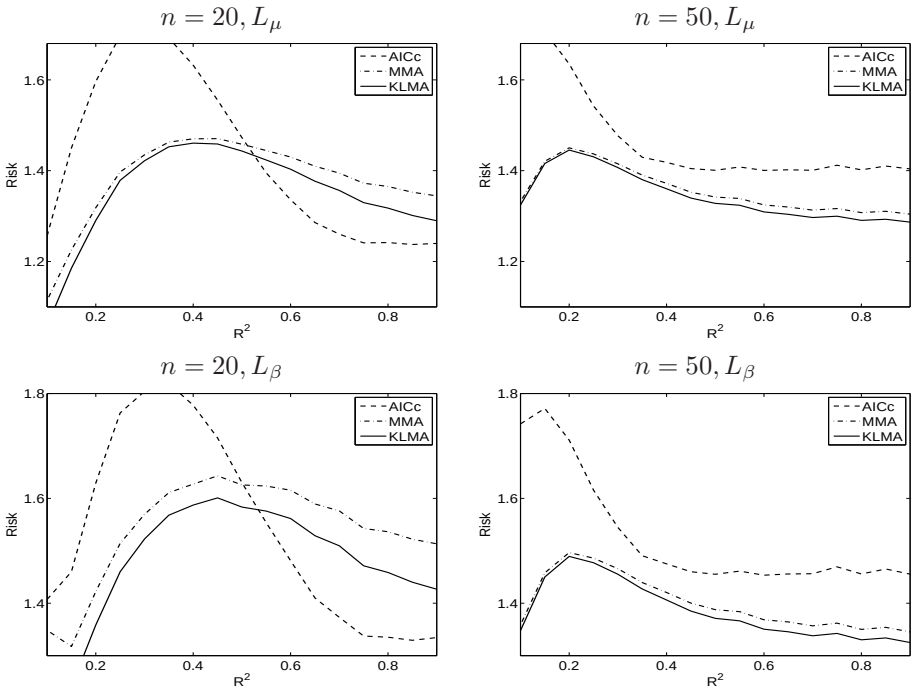

Figure S.1: Results for Example 1: risk comparisons under L_μ and L_β as a function of R^2.

S9 All Simulation Results of Example 2

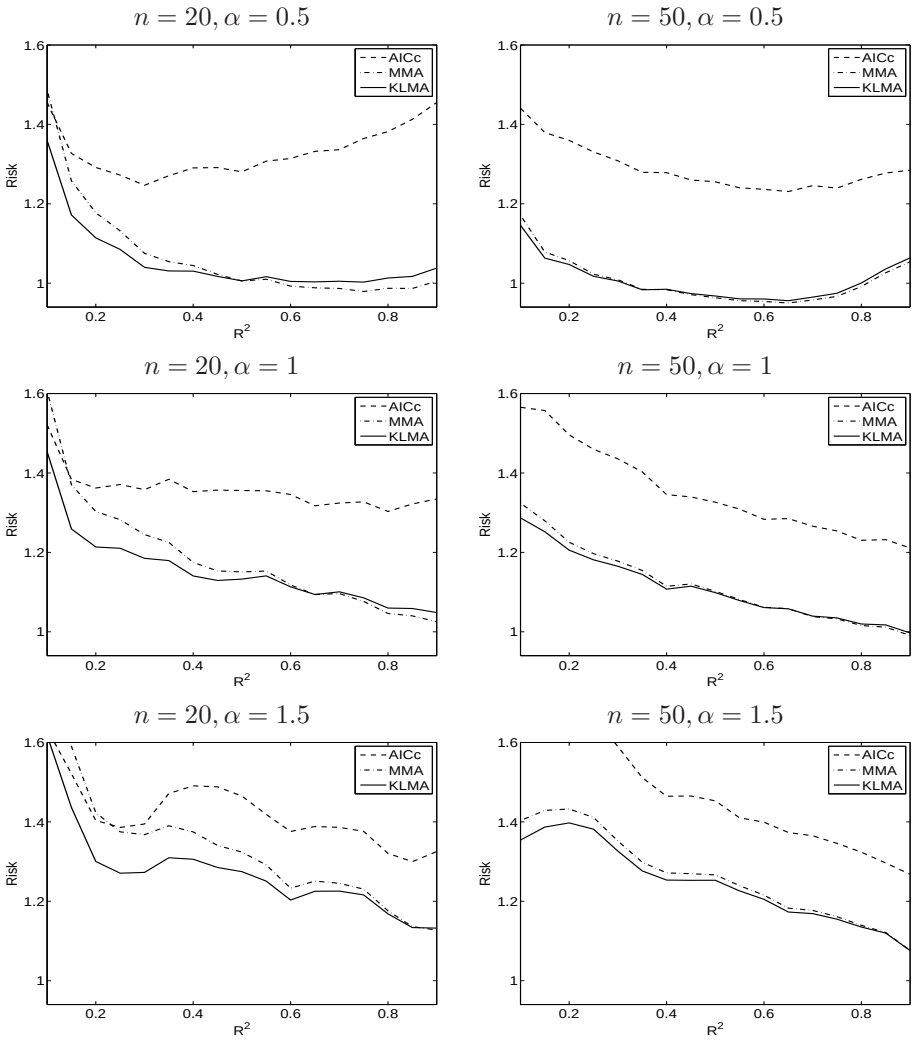

Figure S.2: Results for Example 2: risk comparisons under L_μ as a function of R^2.

S10 All Simulation Results of Example 3

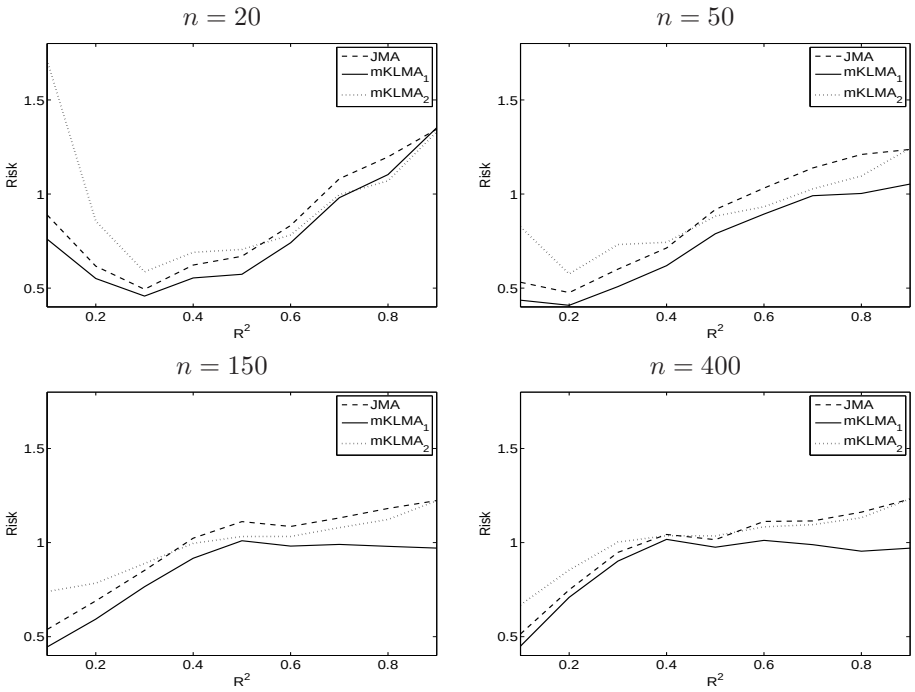

Figure S.3: Results for Example 3: risk comparisons under L_μ as a function of R^2.

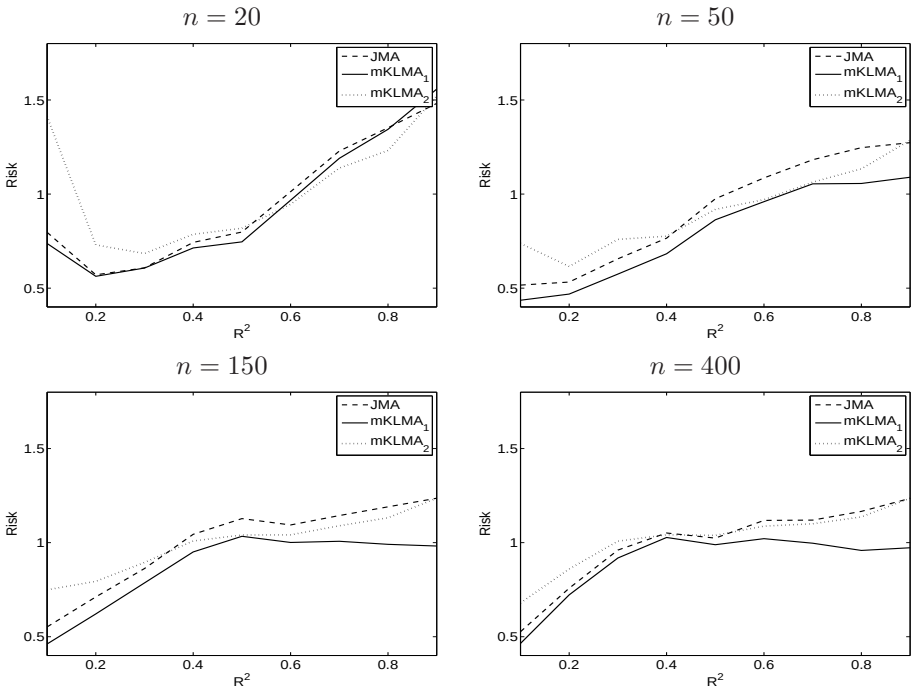

Figure S.4: Results for Example 3: risk comparisons under L_β as a function of R^2.

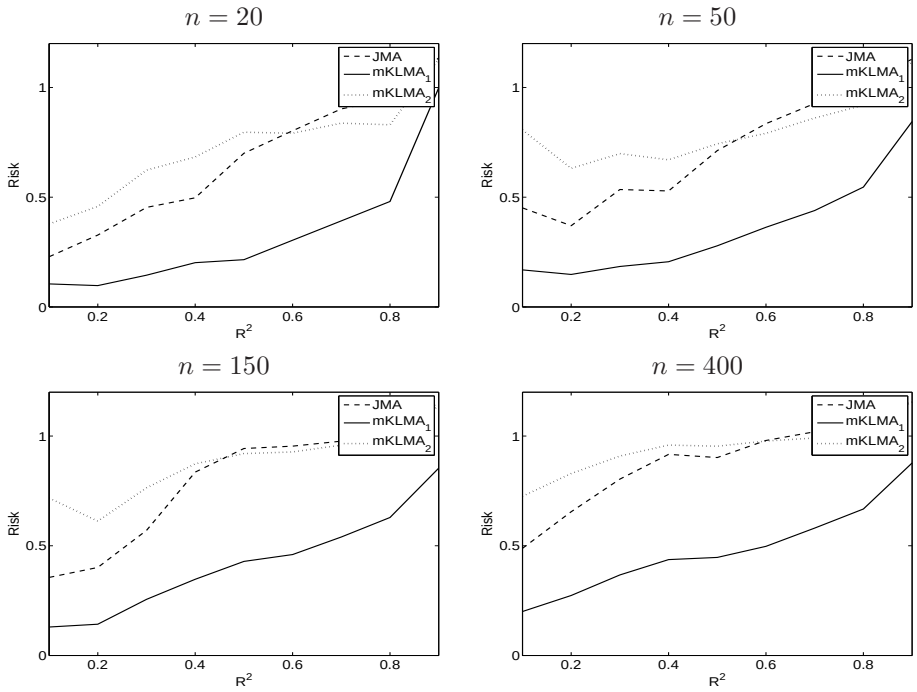

Figure S.5: Results for Example 3: risk comparisons under $L_{\text{hetero},\mu}$ as a function of R^2.

S11 Simulation Results with a Uniformly Distributed Error Term

In this section, we repeat Examples 1-3 but using a uniformly distributed error term.

Example 1	Example 2	Example 3
σUniform$(-3^{1/2}, 3^{1/2})$	Uniform$(-3^{1/2}, 3^{1/2})$	$\exp(\eta X_{2i})^{1/2}$Uniform$(-3^{1/2}, 3^{1/2})$

The following figures show new results.

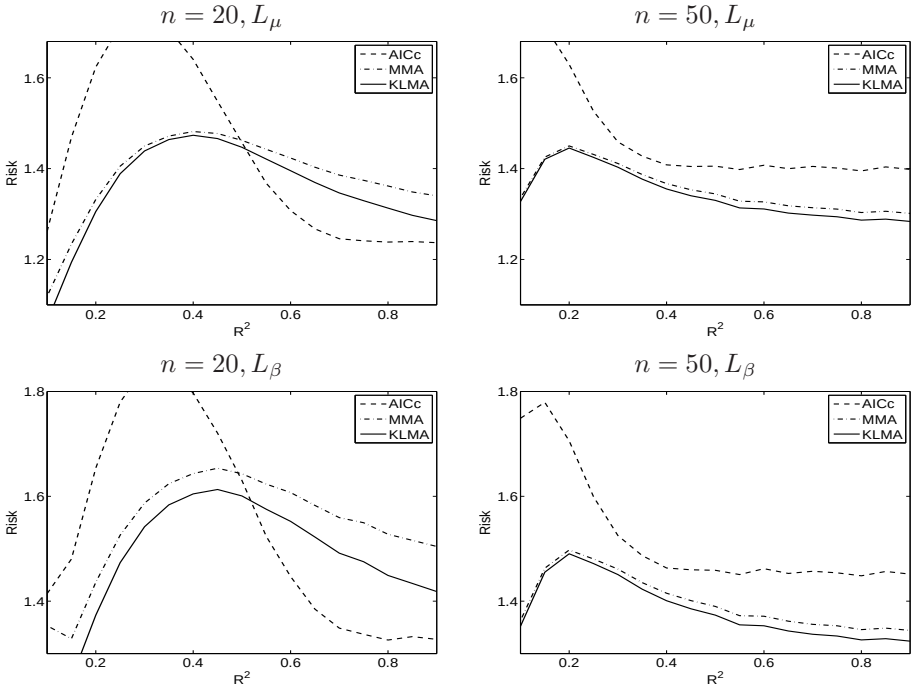

Figure S.6: Results for Example 1 with a uniformly distributed error term: risk comparisons under L_μ and L_β as a function of R^2.

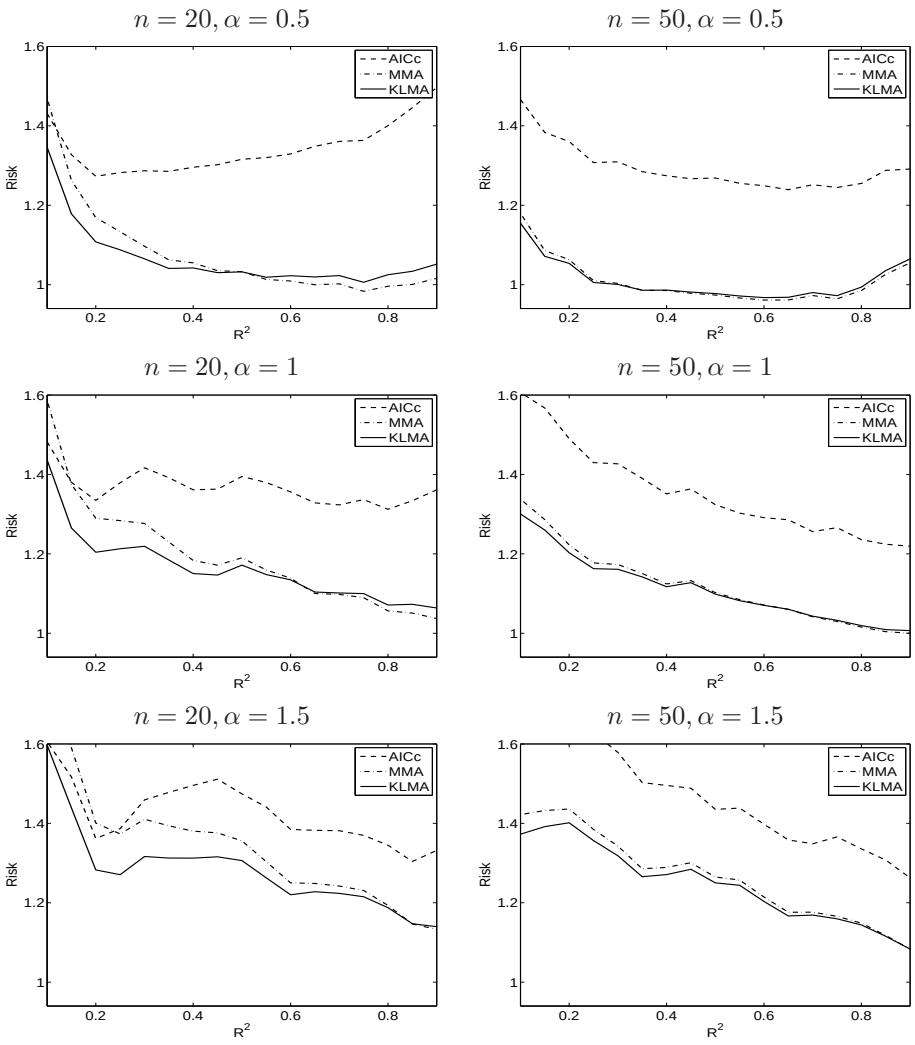

Figure S.7: Results for Example 2 with a uniformly distributed error term: risk comparisons under L_μ as a function of R^2.

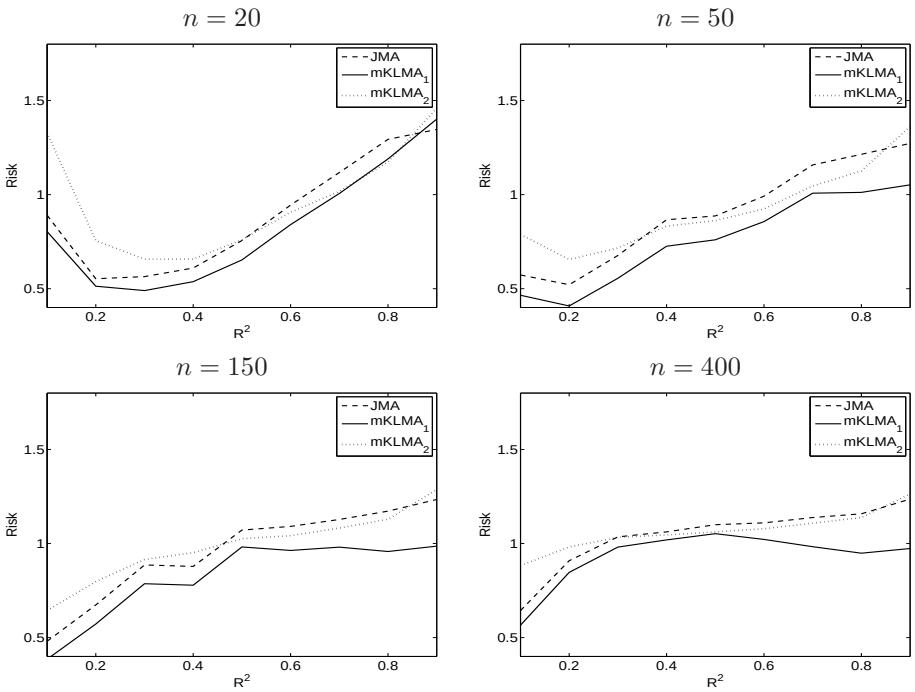

Figure S.8: Results for Example 3 with a uniformly distributed error term: risk comparisons under L_μ as a function of R^2.

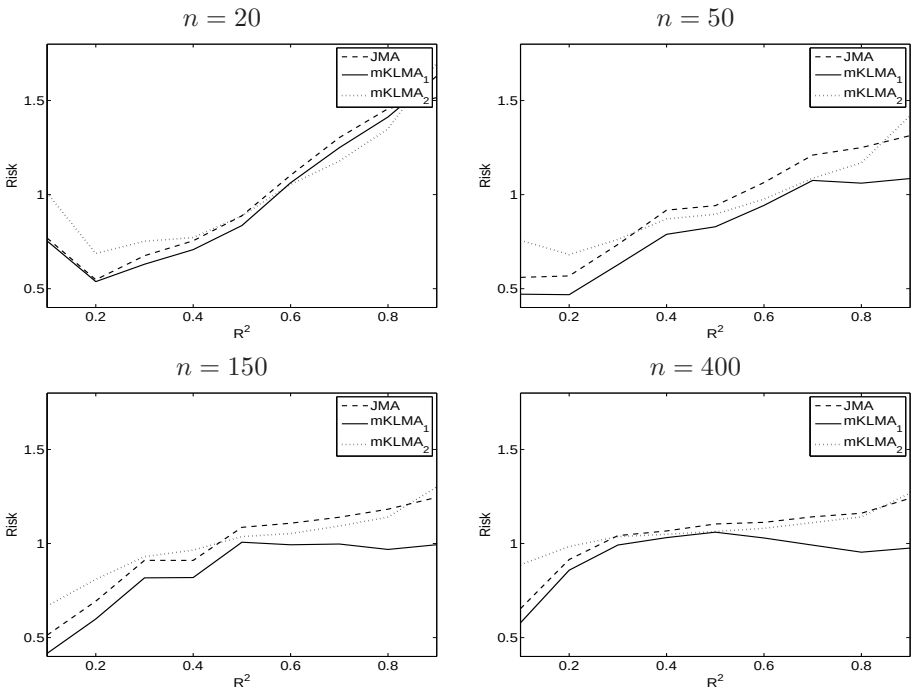

Figure S.9: Results for Example 3 with a uniformly distributed error term: risk comparisons under L_β as a function of R^2.

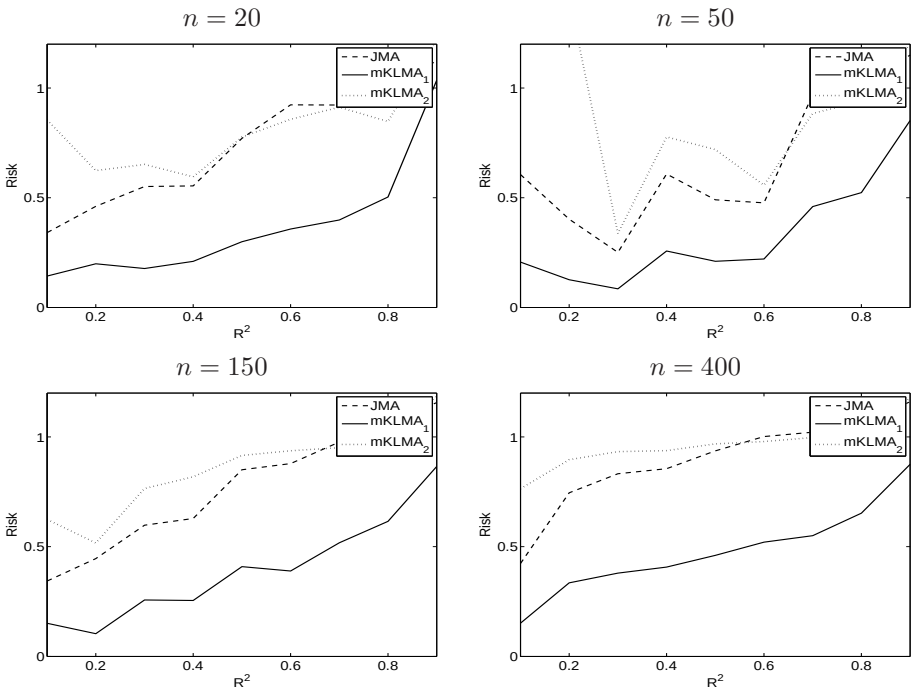

Figure S.10: Results for Example 3 with a uniformly distributed error term: risk comparisons under $L_{\text{hetero},\mu}$ as a function of R^2.

S12 Simulation Results with a Chi-squared Distributed Error Term

In this section, we repeat Examples 1-3 but using a Chi-squared distributed error term:

Example 1	Example 2	Example 3
$\sigma 8^{-1/2}\{\chi^2(4)-4\}$	$8^{-1/2}\{\chi^2(4)-4\}$	$\exp(\eta X_{2i})^{1/2}8^{-1/2}\{\chi^2(4)-4\}$

The following figures show new results.

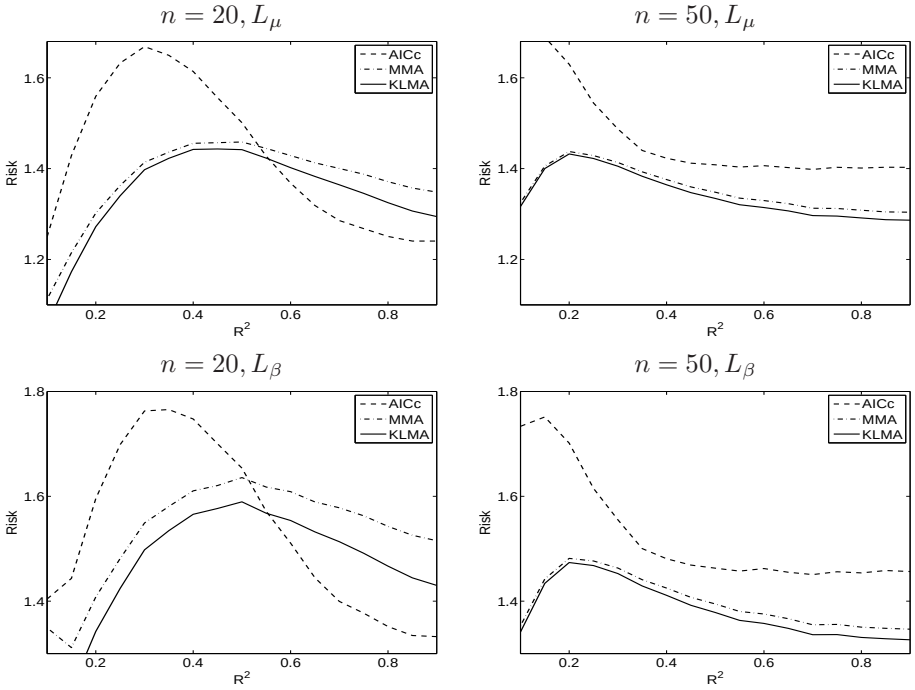

Figure S.11: Results for Example 1 with a Chi-squared distributed error term: risk comparisons under L_μ and L_β as a function of R^2.

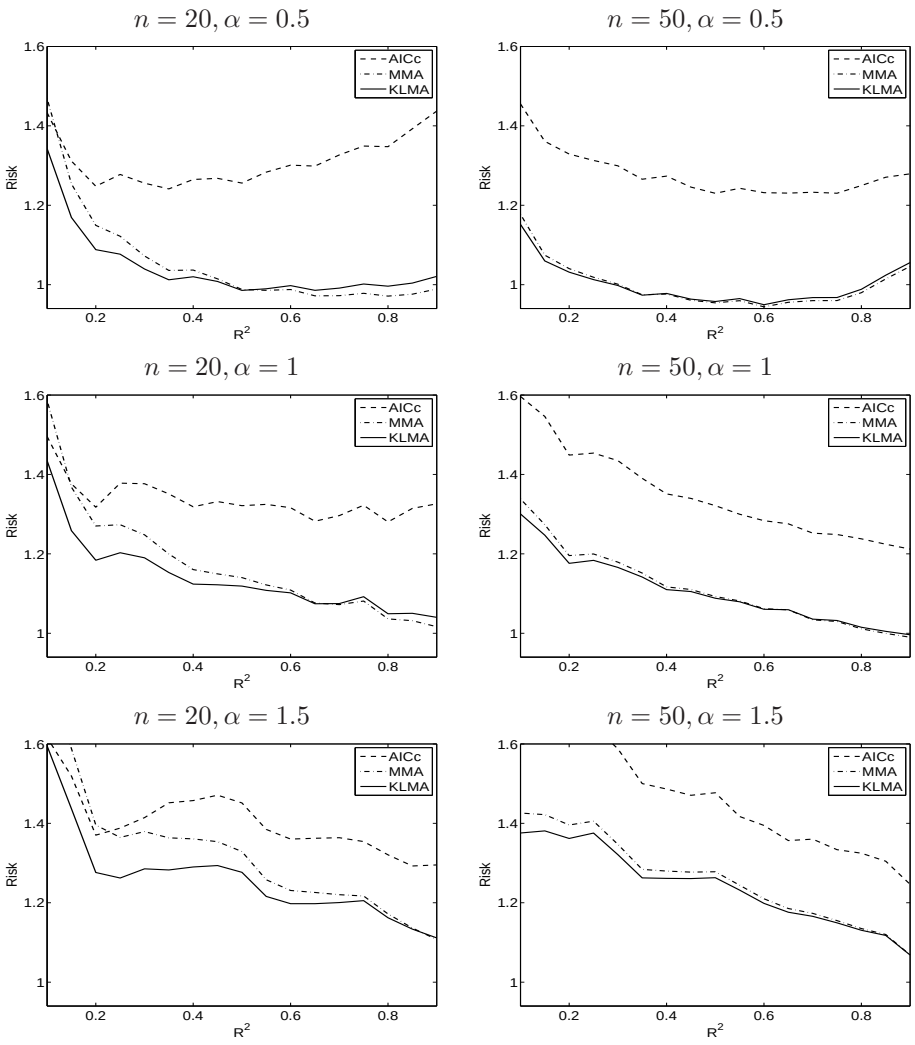

Figure S.12: Results for Example 2 with a Chi-squared distributed error term: risk comparisons under L_μ as a function of R^2.

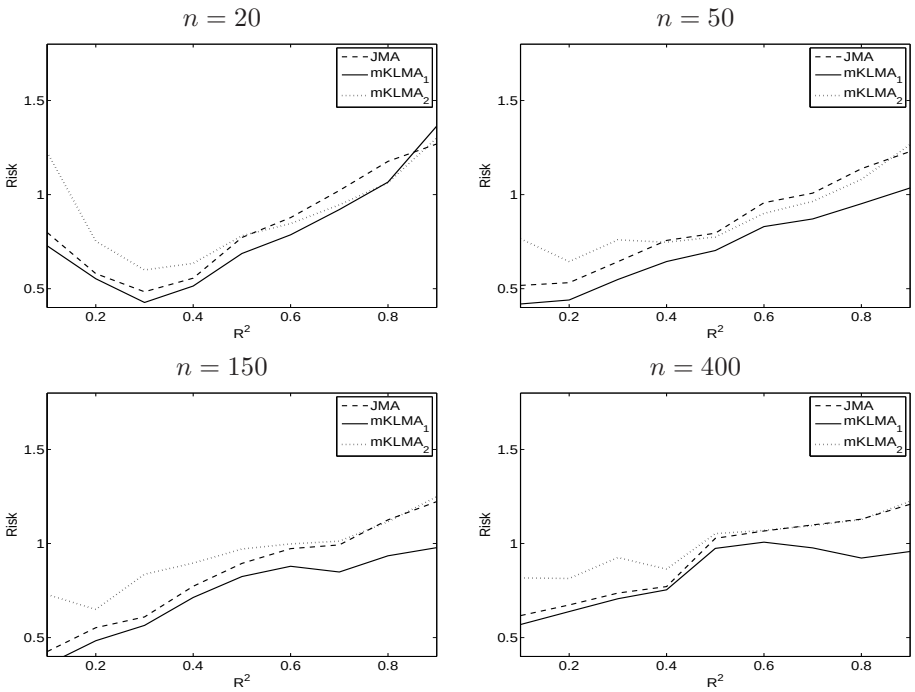

Figure S.13: Results for Example 3 with a Chi-squared distributed error term: risk comparisons under L_μ as a function of R^2.

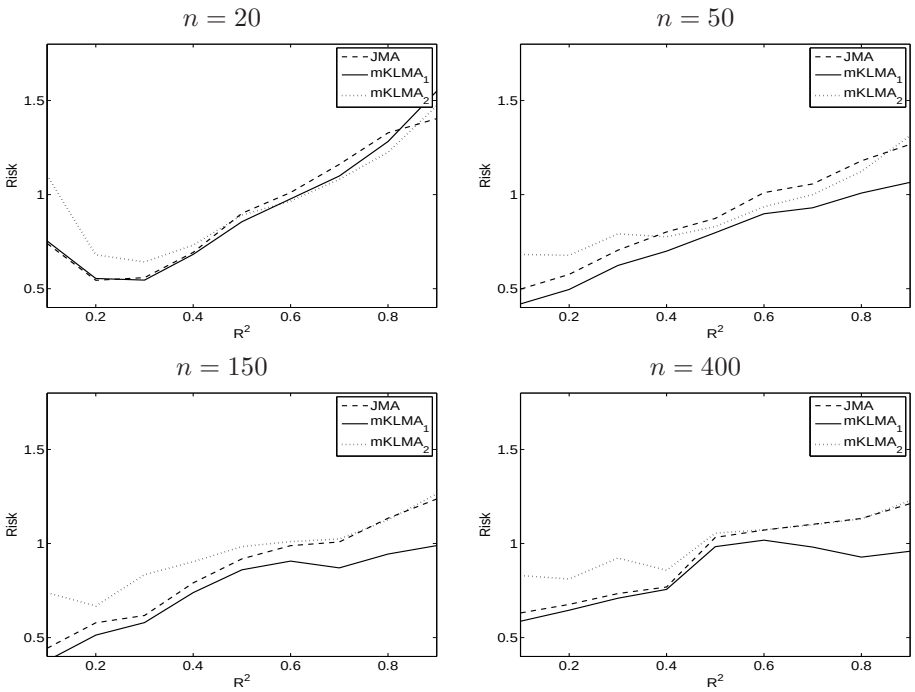

Figure S.14: Results for Example 3 with a Chi-squared distributed error term: risk comparisons under L_β as a function of R^2.

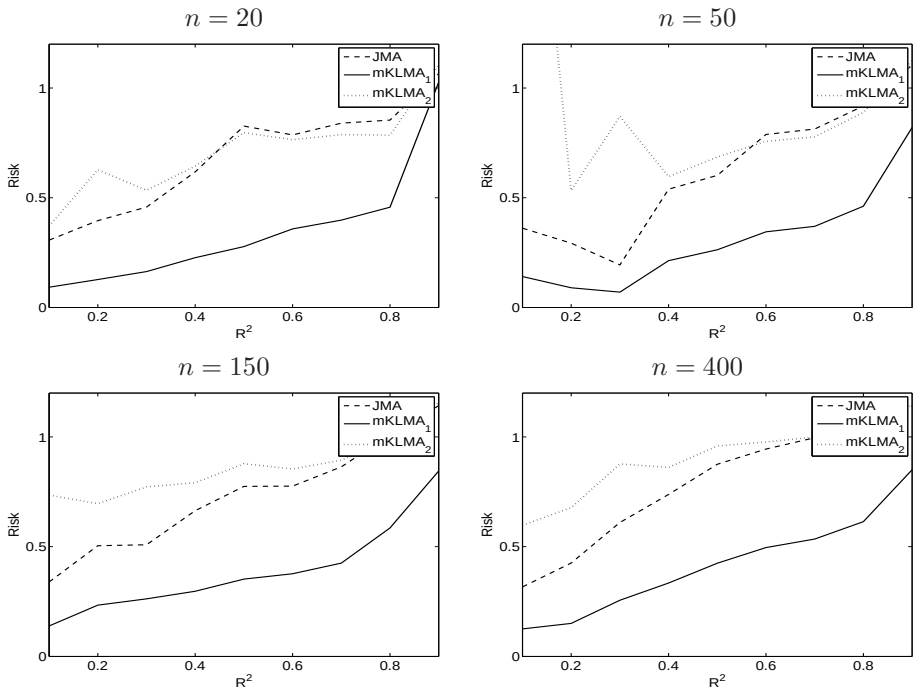

Figure S.15: Results for Example 3 with a Chi-squared distributed error term: risk comparisons under $L_{\text{hetero},\mu}$ as a function of R^2.

S13 Simulation Results with Coefficients Depending on the Sample Size

In this section, we repeat Examples 1 and 3, but let the coefficients depend on the sample size. Specifically, we set $\beta = (1, 2, 3, 2/\sqrt{n}, 2/\sqrt{n}, 2/\sqrt{n}, 2/\sqrt{n})^{\mathrm{T}}$. The following figures show new results.

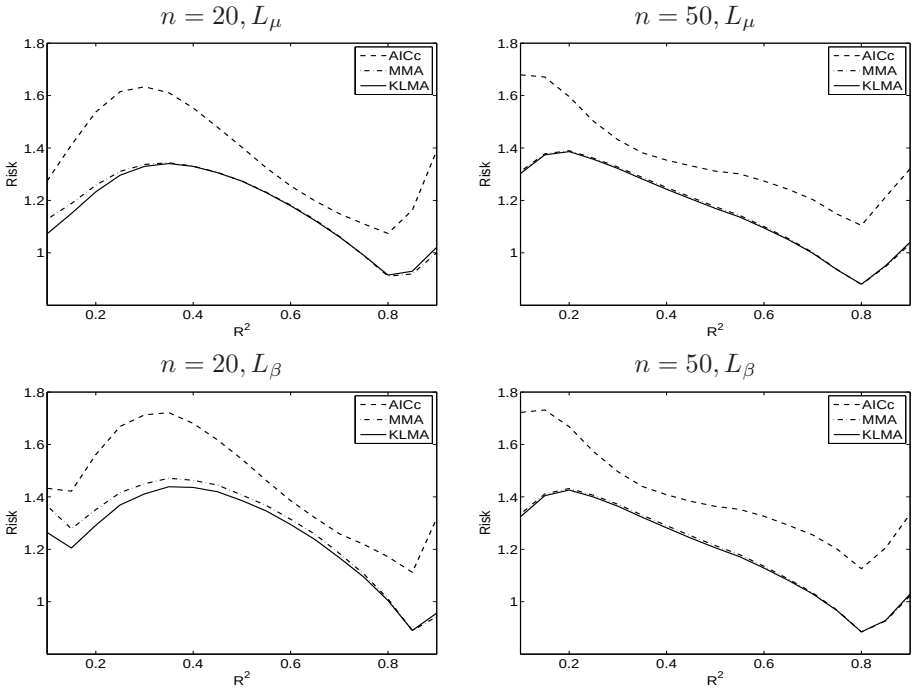

Figure S.16: Results for Example 1 with the coefficients depending on the sample size: risk comparisons under L_μ and L_β as a function of R^2.

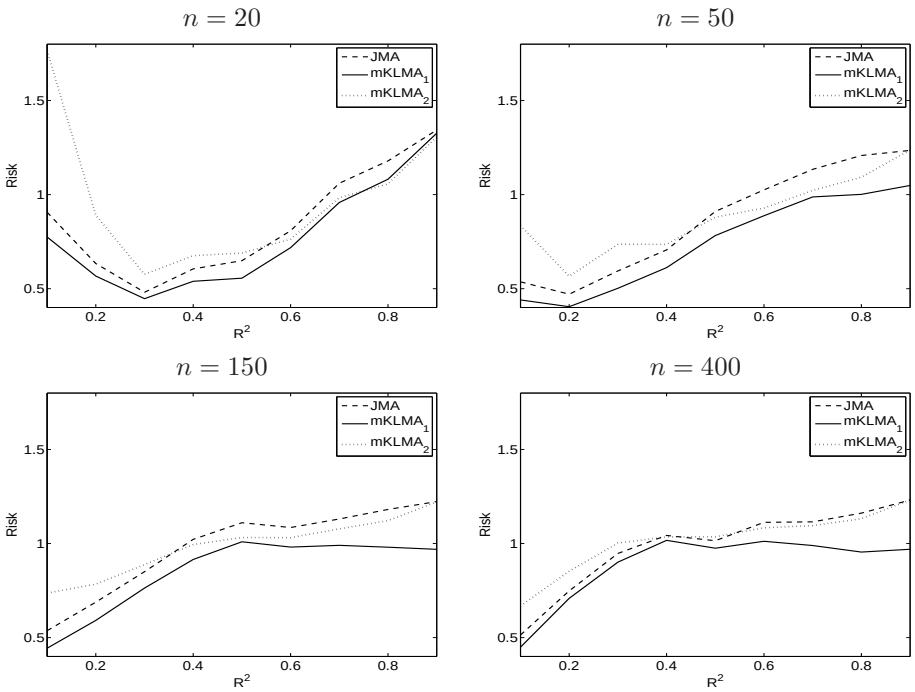

Figure S.17: Results for Example 3 with the coefficients depending on the sample size: risk comparisons under L_μ as a function of R^2.

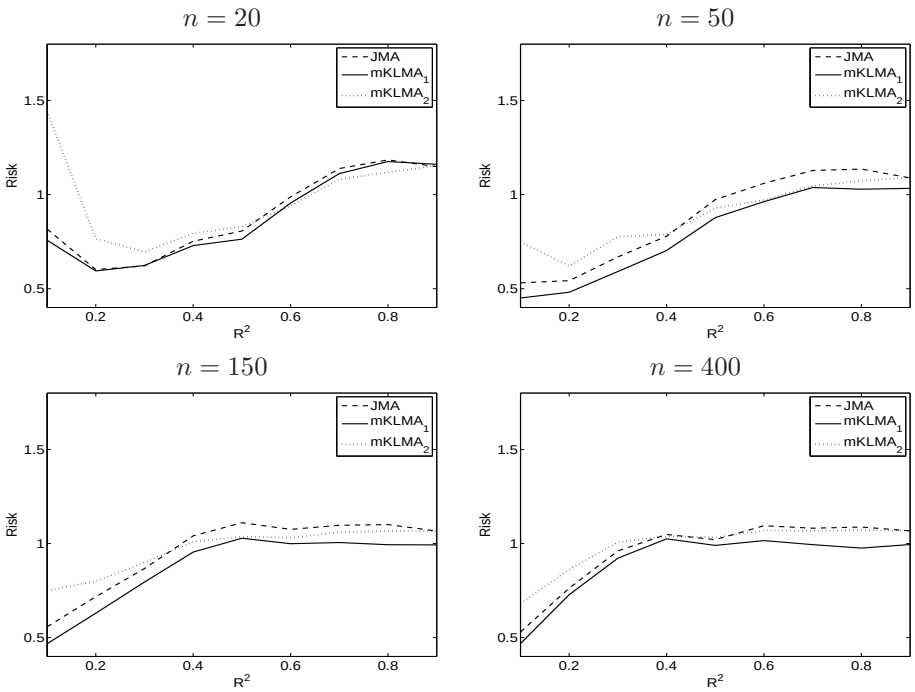

Figure S.18: Results for Example 3 with the coefficients depending on the sample size: risk comparisons under L_β as a function of R^2.

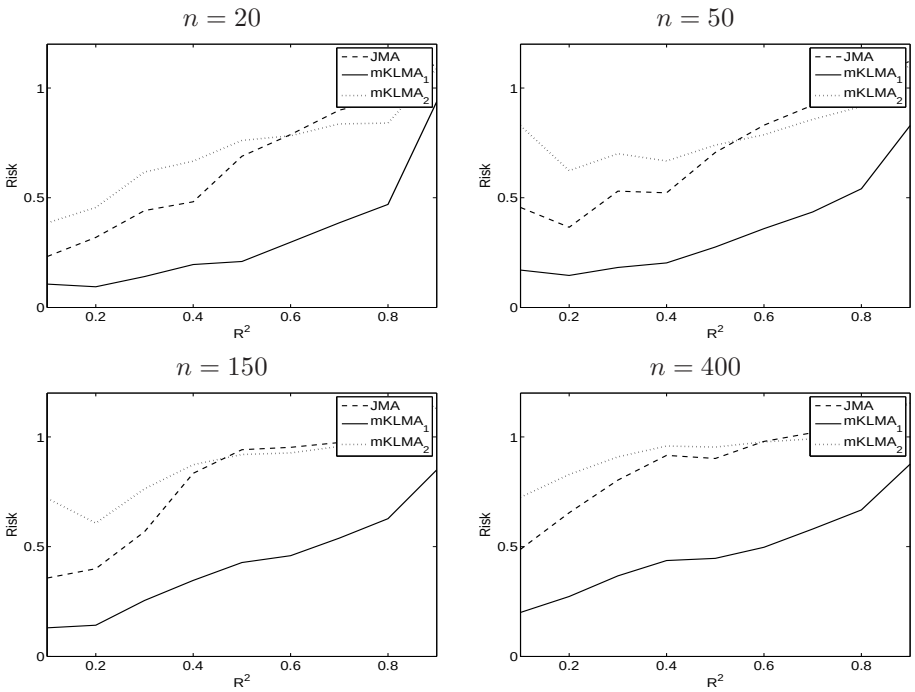

Figure S.19: Results for Example 3 with the coefficients depending on the sample size: risk comparisons under $L_{\text{hetero},\mu}$ as a function of R^2.

References

Hansen, B. E. and Racine, J. (2012). Jacknife model averaging. *Journal of Econometrics* **167**, 38–46.

Liu, Q. and Okui, R. (2013). Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal* **16**, 463–472.

Magnus, J. R and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135–1151.

Wan, A. T. K., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* **156**, 277–283.