Supplementary Materials

Appendix

A SMN distributions and their conditional distribution properties

Assuming $\boldsymbol{X} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, we can generate an n-dimensional SMN random vector $\boldsymbol{Y}$ (denoted by $\boldsymbol{Y} \sim \mathrm{SMN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathrm{H})$) by the transformation

$$\boldsymbol{Y} = \boldsymbol{\mu} + \kappa^{1/2}(r)\boldsymbol{X}, \tag{A.1}$$

where $\boldsymbol{\mu}$ is a location vector, $\kappa(\cdot)$ is a strictly positive weight function, and r is a positive scale random variable (independent of $\boldsymbol{X}$) with its cumulative distribution function $\mathrm{H}(r; \boldsymbol{\nu})$. We use the notation $\boldsymbol{Y} \sim \mathrm{SMN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathrm{H})$. Given r, $\boldsymbol{Y}$ is a multivariate normal distribution, i.e., $\boldsymbol{Y}|r \sim \mathrm{N}_n(\boldsymbol{\mu}, \kappa(r)\boldsymbol{\Sigma})$. Hence, the marginal density function of $\boldsymbol{Y}$ can be expressed as

$$p(\boldsymbol{y}) = \int_0^\infty \phi_n(\boldsymbol{y}; \boldsymbol{\mu}, \kappa(r)\boldsymbol{\Sigma}) \ \mathrm{dH}(r), \tag{A.2}$$

where $\phi_n(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the pdf of the n-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Some SMN distributions and their conditional distribution properties are as follows:

(1) The multivariate Student-t distribution

When $\kappa(r) = 1/r$ and $r \sim \mathrm{Gamma}(\nu/2, \nu/2)$, $\boldsymbol{Y}$ follows a multivariate Student-t distribution $t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu)$, with pdf as

$$p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma((\nu+n)/2)}{\Gamma(\nu/2)(\nu/2)^{n/2}} |2\pi\Sigma|^{-1/2}(1 + d/\nu)^{-(\nu+n)/2}, \tag{A.3}$$

where $d = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ is the Mahalanobis distance. The multi-normal distribution is the limiting case when $\nu \to +\infty$. Given $\boldsymbol{Y} = \boldsymbol{y}$, the conditional distribution of r is $\mathrm{Gamma}(\frac{\nu+n}{2}, \frac{\nu+d}{2})$. It comes the conditional expectation

$$\mathrm{E}[r^m|\boldsymbol{y}] = \frac{2^m\Gamma((\nu+n+2m)/2)(\nu+d)^{-m}}{\Gamma((\nu+n)/2)}.$$

(2) The multivariate slash distribution

When $\kappa(r) = 1/r$ and $r \sim \text{Beta}(\nu, 1)$, we get the multivariate slash distribution $\text{SL}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu))$ with pdf as

$$p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \begin{cases} \nu|2\pi\boldsymbol{\Sigma}|^{-1/2}\Gamma(\nu + n/2)P_1(\nu + n/2, d/2)(d/2)^{-(\nu+n/2)}, & \boldsymbol{y} \neq \boldsymbol{\mu}, \\ |2\pi\boldsymbol{\Sigma}|^{-1/2}\nu/(\nu + n/2), & \boldsymbol{y} = \boldsymbol{\mu}, \end{cases} \tag{A.4}$$

where $P_x(a, b)$ denotes the cumulative distribution function of the Gamma(a, b) distribution. When $\nu \to +\infty$, the slash distribution reduces to the normal distribution. The conditional distribution of r given $\boldsymbol{y}$ is a truncated gamma distribution $\text{Gamma}(\nu + n/2, d/2)\mathbb{I}_{(0,1)}$. Then, we get

$$\text{E}[r^m|\boldsymbol{y}] = \frac{\Gamma(\nu + n/2 + m)}{\Gamma(\nu + n/2)}(d/2)^{-m}\frac{P_1(\nu + n/2 + m, d/2)}{P_1(\nu + n/2, d/2)}.$$

(3) The contaminated-normal distribution

When $\kappa(r) = 1/r$ and r is a discrete random variable with pdf $h(r; \nu, \gamma) = \nu\mathbb{I}_{(r=\gamma)} + (1 - \nu)\mathbb{I}_{(r=1)}$, with $0 < \nu \leqslant 1, 0 < \gamma \leqslant 1$, we obtain the multivariate contaminated-normal distribution $\text{CN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu, \gamma)$. Its pdf is given by

$$p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu, \gamma) = \nu\phi_n(\boldsymbol{y}; \boldsymbol{\mu}, \gamma^{-1}\boldsymbol{\Sigma}) + (1 - \nu)\phi_n(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A.5}$$

When $\gamma = 1$, it reduces to the normal distribution. Given $\boldsymbol{y}$, r is a discrete random variable with the conditional distribution as $\widetilde{h}(r; \tilde{\nu}, \gamma) = \tilde{\nu}\mathbb{I}_{(r=\gamma)} + (1 - \tilde{\nu})\mathbb{I}_{(r=1)}$, with $1/\tilde{\nu} = 1 + (1/\nu - 1)\gamma^{-n/2}\exp(-\frac{1-\gamma}{2}d)$. Hence, we get $\text{E}[r^m|\boldsymbol{y}] = \tilde{\nu}\gamma^m + 1 - \tilde{\nu}$.

B The observed and expected information matrix

We provide the information matrix of $\boldsymbol{\Theta}$ for the HPFR model. Since the SMN distributions belong to the elliptical distributions class (Fang et al., 1990), the observed response $\boldsymbol{y}_m$ of the HPFR model follows an elliptical distribution $\text{EL}_{n_m}(\widetilde{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m; g_m)$, where $g_m(\cdot) : \mathbb{R} \to [0, \infty)$ is the density generator such that $\int_0^\infty g_m(u; \boldsymbol{\nu})du < \infty$. The pdf of $\boldsymbol{y}_m$ is given by

$$p(\boldsymbol{y}_m) = \left|\boldsymbol{\Sigma}_m\right|^{-1/2} g_m(d_m; \boldsymbol{\nu}), \ m = 1, \ldots, M,$$

where $d_m = (\boldsymbol{y}_m - \boldsymbol{\mu}_m)^\top \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{y}_m - \boldsymbol{\mu}_m)$, and

$$g_m(d_m; \boldsymbol{\nu}) = (2\pi)^{-n_m/2} \int_0^\infty \kappa^{-n_m/2}(r) \exp\{-\kappa^{-1}(r)d_m/2\} \, \text{dH}(r; \boldsymbol{\nu}).$$

Thus, the log-likelihood function for $\boldsymbol{\Theta}$ is given by

$$l(\boldsymbol{\Theta}) = \sum_{m=1}^{M} l_m(\boldsymbol{\Theta}) = -\frac{1}{2}\sum_{m=1}^{M} \log(\left|\boldsymbol{\Sigma}_m\right|) + \sum_{m=1}^{M} \log\{g_m(d_m;\boldsymbol{\nu})\}, \tag{B.1}$$

and the score function of $\boldsymbol{\Theta}$ has a form as

$$\frac{\partial}{\partial\Theta_i} l(\boldsymbol{\Theta}) = -\frac{1}{2}\sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\Theta_i}) + \sum_{m=1}^{M} \dot{g}_{m,\Theta_i}/g_m, \tag{B.2}$$

where $\dot{\boldsymbol{\Sigma}}_{m,\Theta_i}$ and $\dot{g}_{m,\Theta_i}$ mean respectively, $\partial\boldsymbol{\Sigma}_m/\partial\Theta_i$ and $\partial g_m/\partial\Theta_i$.

Denoting

$$I_m(\omega) = (2\pi)^{-n_m/2}\int_0^\infty \kappa^{-\omega}(r)\exp\{-\kappa^{-1}(r)d_m/2\}\,\mathrm{dH}(r;\boldsymbol{\nu}), \omega > 0, \tag{B.3}$$

then g_m and $\dot{g}_{m,\Theta_i}$ (with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$) can be expressed respectively as $I_m(n_m/2)$ and $-I_m(n_m/2+1)\dot{d}_{m,\Theta_i}/2$. We can find g_m for some SMN distributions in Appendix A. Specific forms of $I_m(\omega)$ and $\partial\log(g_m)/\partial\boldsymbol{\nu}$ or $\partial g_m/\partial\boldsymbol{\nu}$ are given below,

(1) for Student-t:

$$\begin{aligned} I_m(\omega) &= (2\pi)^{-n_m/2}2^{\omega}\nu^{\nu/2}\Gamma(\nu/2+\omega)/\Gamma(\nu/2)(d_m+\nu)^{-(\nu/2+\omega)}, \\ \frac{\partial\log(g_m)}{\partial\nu} &= \frac{1}{2}\varphi(\frac{\nu+n_m}{2}) - \frac{1}{2}\varphi(\frac{\nu}{2}) - \frac{1}{2}\log(1+\frac{d_m}{\nu}) + \frac{d_m-n_m}{2(\nu+d_m)}, \end{aligned}$$

where $\varphi(x) = \mathrm{d}\log(\Gamma(x))/\mathrm{d}x$ is the digamma function.

(2) for slash:

$$\begin{aligned} I_m(\omega) &= (2\pi)^{-n_m/2}2^{\nu+\omega}\nu\Gamma(\nu+\omega)P_1(\nu+\omega, d_m/2)d_m^{-(\nu+\omega)}, \\ \frac{\partial\log(g_m)}{\partial\nu} &= 1/\nu + c_m, \end{aligned}$$

where $c_m = \mathrm{E}[\log(X)]$ and X follows a truncated gamma distribution $\mathrm{Gamma}(\nu + n_m/2, d_m/2)I(0,1)$.

(3) for contaminated-normal:

$$\begin{aligned} I_m(\omega) &= (2\pi)^{-(n_m-1)/2}[\nu\gamma^{\omega}\phi_1(\sqrt{\gamma d_m}) + (1-\nu)\phi_1(\sqrt{d_m})], \\ \frac{\partial g_m}{\partial\nu} &= (2\pi)^{-(n_m-1)/2}[\gamma^{n_m/2}\phi_1(\sqrt{\gamma d_m}) - \phi_1(\sqrt{d_m})], \\ \frac{\partial g_m}{\partial\gamma} &= (2\pi)^{-(n_m-1)/2}\nu\gamma^{n_m/2-1}(n_m-\gamma d_m)\phi_1(\sqrt{\gamma d_m})/2. \end{aligned}$$

The observed information matrix $\mathbf{J}(\widehat{\boldsymbol{\Theta}})$ can be approximated by $\sum_{m=1}^{M}\widehat{\boldsymbol{s}}_m\widehat{\boldsymbol{s}}_m^\top$ (McLachlan and Basford, 1988), where $\widehat{\boldsymbol{s}}_m = \partial l_m(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}|_{\widehat{\boldsymbol{\Theta}}}$. By calculating the expectation of the second-order derivatives of (B.1), we can obtain the Fisher information matrix $\mathbf{I}(\boldsymbol{\Theta}) = (\mathrm{I}_{\Theta_i\Theta_j})_{p\times p}$, in which p is the dimension of $\boldsymbol{\Theta}$. The elements of the information matrix are calculated by

$$\begin{aligned}
\mathrm{I}_{\beta_i\beta_j} &= \sum_{m=1}^{M} \frac{4}{n_m} d_{g,m} \dot{\tilde{\boldsymbol{\mu}}}_{m,\beta_i}^{\top} \boldsymbol{\Sigma}_m^{-1} \dot{\tilde{\boldsymbol{\mu}}}_{m,\beta_j},\\
\mathrm{I}_{\psi_i\psi_j} &= \sum_{m=1}^{M} \Big[a_m \mathrm{tr}(\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\psi_i}\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\psi_j}) + b_m \mathrm{tr}(\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\psi_i})\mathrm{tr}(\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\psi_j})\Big],\\
\mathrm{I}_{\psi_i\nu_j} &= \sum_{m=1}^{M} \frac{1}{n_m} \mathrm{E}[d_m \frac{\partial}{\partial \nu_j}(W_{g_m})]\mathrm{tr}(\boldsymbol{\Sigma}_m^{-1}\dot{\boldsymbol{\Sigma}}_{m,\psi_i}),\\
\mathrm{I}_{\nu_i\nu_j} &= -\sum_{m=1}^{M} \mathrm{E}[\frac{\partial^2}{\partial\nu_i\partial\nu_j}\log(g_m)],\\
\mathrm{I}_{\beta_i\psi_j} &= \mathrm{I}_{\beta_i\nu_j} = 0,
\end{aligned}$$

where $a_m = \frac{2f_{g,m}}{n_m(n_m+2)}$, $b_m = \frac{f_{g,m}}{n_m(n_m+2)} - \frac{1}{4}$, $f_{g,m} = \mathrm{E}(W_{g_m}^2 d_m^2)$, $d_{g,m} = \mathrm{E}(W_{g_m}^2 d_m)$, in which $W_{g_m} = \frac{\partial \log(g_m)}{\partial d_m}$ with $d_m = \boldsymbol{e}_m^{\top}\boldsymbol{e}_m$ and $\boldsymbol{e}_m \sim \mathrm{EL}_{n_m}(\mathbf{0}, \boldsymbol{I}_{n_m}; g_m)$. The asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated via $\mathbf{I}^{-1}(\widehat{\boldsymbol{\Theta}})$. The expectation values of $f_{g,m}$ and $d_{g,m}$ for some SMN distributions (e.g., normal, Student-t and slash) have closed forms (Cao et al., 2015). For contaminated-normal and other distributions, we need to use numerical integration or Monte Carlo approximation.

C Technical details for information consistency

Lemma 1 *Suppose $\boldsymbol{y}_n$ are generated from model (1) with $\tau_0 \in \mathcal{F}$ and we fit them by SMGP with bounded covariance kernel function $C(\cdot,\cdot;\boldsymbol{\theta})$ for any covariate values in $\mathcal{X}$. Suppose $C(\cdot,\cdot;\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely as $n \to \infty$. Then we have*

$$-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|\boldsymbol{X}_n) + \log p(\boldsymbol{y}_n|\tau_0, \boldsymbol{X}_n) \leqslant \frac{1}{2}\{c + \log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n| + b(\|\tau_0\|_c^2 + c)\}, \quad \text{(C.1)}$$

where $\|\tau_0\|_c$ is the reproducing kernel Hilbert space (RKHS) norm of τ_0 associated with $C(\cdot,\cdot;\boldsymbol{\theta})$ and $\boldsymbol{C}_n = (C(\boldsymbol{x}_i, \boldsymbol{x}_j))_{n\times n}$, ϕ, b and c are some positive constants.

Proof. From the hierarchical structure of SMGP, we can rewrite the HPFR model (omit subscript m) conditional on r by

$$y(t) = \mu(t) + \breve{\tau}(t) + \breve{\varepsilon}(t), \quad \text{(C.2)}$$

where $\breve{\tau} = \tau|r \sim \mathrm{GP}(0, \kappa(r)C(\cdot,\cdot;\boldsymbol{\theta}))$ which is independent with the error term $\breve{\varepsilon} = \varepsilon|r \sim \mathrm{N}(0, \kappa(r)\phi)$.

Let

$$p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n) = \int_{\mathcal{F}} p(\boldsymbol{y}_n|r, \breve{\tau}, \boldsymbol{X}_n) \ \mathrm{d}p_{\widehat{\boldsymbol{\theta}}}(\breve{\tau}), \tag{C.3}$$

where $p_{\widehat{\boldsymbol{\theta}}}(\breve{\tau})$ is the induced measure from $\mathrm{GP}(0, \kappa(r)C(\cdot,\cdot;\widehat{\boldsymbol{\theta}}))$. Then we have

$$p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|\boldsymbol{X}_n) = \int p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n)h(r) \ \mathrm{d}r \tag{C.4}$$

and

$$p(\boldsymbol{y}_n|\tau_0, \boldsymbol{X}_n) = \int p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n)h(r) \ \mathrm{d}r. \tag{C.5}$$

Let $\mathcal{H}$ be the RKHS associated with covariance kernel function $C(\cdot,\cdot;\boldsymbol{\theta})$, and $\mathcal{H}_n$ be the span of $\{C(\cdot, \boldsymbol{x}_i;\boldsymbol{\theta})|i = 1,\dots,n\}$, i.e., $\mathcal{H}_n = \{\breve{f}(\cdot): \ \breve{f}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i C(\boldsymbol{x}, \boldsymbol{x}_i;\boldsymbol{\theta})$, for any $\alpha_i \in \mathbb{R}\}$. Assuming the true underlying function $\breve{\tau}_0 = \tau_0|r \in \mathcal{H}_n$, then given r, $\tau_0(\cdot)$ can be expressed as

$$\tau_0(\cdot) = \kappa(r)\sum_{i=1}^n \alpha_i C(\cdot, \boldsymbol{x}_i;\boldsymbol{\theta}) \triangleq \kappa(r)\boldsymbol{C}(\cdot)\boldsymbol{\alpha},$$

where $\boldsymbol{C}(\cdot) = (C(\cdot, \boldsymbol{x}_1;\boldsymbol{\theta}),\dots,C(\cdot, \boldsymbol{x}_n;\boldsymbol{\theta}))$ and $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_n)^\top$.

By Fenchel-Legendre duality relationship, we have

$$-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n) \leqslant \mathrm{E}_{\bar{P}}[-\log p(\boldsymbol{y}_n|r, \breve{\tau}, \boldsymbol{X}_n)] + \mathrm{D}[\bar{P}, P], \tag{C.6}$$

where P is the measure induced by $\mathrm{GP}(0, \kappa(r)C(\cdot,\cdot;\widehat{\boldsymbol{\theta}}))$, and $\bar{P}$ is the posterior distribution of $\breve{\tau}$ from a GP model with prior $\mathrm{GP}(0, \kappa(r)C(\cdot,\cdot;\boldsymbol{\theta}))$ and Gaussian likelihood term $\prod_{i=1}^n \mathrm{N}(\widehat{\boldsymbol{y}}_n|\breve{\tau}(\boldsymbol{x}_i), \kappa(r)\phi)$, where $\widehat{\boldsymbol{y}}_n = \kappa(r)(\boldsymbol{C}_n + \phi\boldsymbol{I}_n)\boldsymbol{\alpha}$ and $\phi > 0$ is a constant to be specified. Then we have

$$\begin{aligned}\mathrm{D}[\bar{P}, P] =& \frac{1}{2}\{-\log|\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n| + \log|\boldsymbol{B}_n| + \mathrm{tr}(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n\boldsymbol{B}_n^{-1}) \\ &+ \kappa(r)\|\tau_0\|_c^2 + \kappa(r)\boldsymbol{\alpha}^\top\boldsymbol{C}_n(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n - \boldsymbol{I}_n)\boldsymbol{\alpha} - n\},\end{aligned} \tag{C.7}$$

and

$$\mathrm{E}_{\bar{P}}[-\log p(\boldsymbol{y}_n|r, \breve{\tau}, \boldsymbol{X}_n)] \leqslant -\log p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n) + \frac{\delta}{2}\mathrm{tr}(\boldsymbol{C}_n\boldsymbol{B}_n^{-1}), \tag{C.8}$$

where $\boldsymbol{B}_n = \boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n$, $\widehat{\boldsymbol{C}}_n$ is the estimation of $\boldsymbol{C}_n$ at $\widehat{\boldsymbol{\theta}}$ and δ is a generic positive constant. Combining (C.6)-(C.8) gives

$$\begin{aligned}&-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n) + \log p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n) \\ \leqslant& \frac{1}{2}\{-\log|\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n| + \log|\boldsymbol{B}_n| + \mathrm{tr}(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n\boldsymbol{B}_n^{-1} + \delta\boldsymbol{C}_n\boldsymbol{B}_n^{-1}) \\ &+ \kappa(r)\|\tau_0\|_c^2 + \kappa(r)\boldsymbol{\alpha}^\top\boldsymbol{C}_n(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n - \boldsymbol{I}_n)\boldsymbol{\alpha} - n\}.\end{aligned} \tag{C.9}$$

Since the covariance function is bounded and continuous in $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$, we have $\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n - \boldsymbol{I}_n \to \boldsymbol{0}$ as $n \to \infty$. Hence, there exist some positive constant c and ε such that

$$\begin{aligned} &-\log|\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n| < c, \quad \boldsymbol{\alpha}^\top\boldsymbol{C}_n(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n - \boldsymbol{I}_n)\boldsymbol{\alpha} < c, \\ &\mathrm{tr}(\widehat{\boldsymbol{C}}_n^{-1}\boldsymbol{C}_n\boldsymbol{B}_n^{-1}) < \mathrm{tr}((\boldsymbol{I}_n + \varepsilon\boldsymbol{C}_n)\boldsymbol{B}_n^{-1}). \end{aligned} \tag{C.10}$$

Thus we have

$$\begin{aligned} &-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n) + \log p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n) \\ \leqslant&\frac{1}{2}\{c + \log|\boldsymbol{B}_n| + \mathrm{tr}((\boldsymbol{I}_n + (\varepsilon+\delta)\boldsymbol{C}_n)\boldsymbol{B}_n^{-1}) \\ &+ \kappa(r)(\|\tau_0\|_c^2 + c) - n\}. \end{aligned} \tag{C.11}$$

Letting $\phi = 1/(\varepsilon + \delta)$, we get

$$\begin{aligned} &-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|r, \boldsymbol{X}_n) + \log p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n) \\ \leqslant&\frac{1}{2}\{c + \log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n| + \kappa(r)(\|\tau_0\|_c^2 + c)\}. \end{aligned} \tag{C.12}$$

It follows that

$$\begin{aligned} -\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|\boldsymbol{X}_n) \leqslant&\frac{1}{2}\{c + \log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n|\} \\ &-\log\int p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n)\exp\{-\frac{1}{2}\kappa(r)(\|\tau_0\|_c^2 + c)\}h(r)\ \mathrm{d}r. \end{aligned} \tag{C.13}$$

Denote $\widetilde{h}(r) \triangleq p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n)h(r)/p(\boldsymbol{y}_n|\tau_0, \boldsymbol{X}_n)$ be the conditional density function of r given $\boldsymbol{y}_n$ and τ_0, then we have

$$\begin{aligned} &\int p(\boldsymbol{y}_n|r, \tau_0, \boldsymbol{X}_n)\exp\{-\frac{1}{2}\kappa(r)(\|\tau_0\|_c^2 + c)\}h(r)\ \mathrm{d}r \\ =&p(\boldsymbol{y}_n|\tau_0, \boldsymbol{X}_n)\int\exp\{-\frac{1}{2}\kappa(r)(\|\tau_0\|_c^2 + c)\}\widetilde{h}(r)\ \mathrm{d}r. \end{aligned} \tag{C.14}$$

Plugging (C.14) in (C.13), we get

$$\begin{aligned} &-\log p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|\boldsymbol{X}_n) + \log p(\boldsymbol{y}_n|\tau_0, \boldsymbol{X}_n) \\ \leqslant&\frac{1}{2}\{c + \log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n|\} - \log\int\exp\{-\frac{1}{2}\kappa(r)(\|\tau_0\|_c^2 + c)\}\widetilde{h}(r)\mathrm{d}r \\ \leqslant&\frac{1}{2}\{c + \log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n| + (\|\tau_0\|_c^2 + c)\mathrm{E}[\kappa(r)|\boldsymbol{y}_n, \tau_0]\}, \end{aligned} \tag{C.15}$$

where $\mathrm{E}[\kappa(r)|\boldsymbol{y}_n, \tau_0] = \int\kappa(r)\widetilde{h}(r)\ \mathrm{d}r$. Supposing $\mathrm{E}[\kappa(r)|\boldsymbol{y}_n, \tau_0]$ is bounded, i.e., there exists a positive constant b such that

$$\mathrm{E}[\kappa(r)|\boldsymbol{y}_n, \tau_0] < b, \tag{C.16}$$

taking infimum of the right hand side of (C.15) over τ_0 and applying the Representer Theorem (Seeger et al., 2008), we complete the proof of Lemma 1.

Remark 1 *Lemma 1 requires that* $\mathrm{E}[\kappa(r)|\boldsymbol{y}_n, \tau_0]$ *is bounded* (C.16). *We now prove it is satisfied for some members of SMN distributions.*

(1) For normal distribution:

It is easy to see since $\kappa(r) \equiv 1$.

(2) For Student-t distribution:

Given $\boldsymbol{y}_n$ *and* τ_0*, the conditional distribution of* r *is* $\mathrm{Gamma}(\frac{\nu+n}{2}, \frac{\nu+d_n}{2})$ *with* $d_n = (\boldsymbol{y}_n - \boldsymbol{\tau}_0(\boldsymbol{X}_n))^\top(\boldsymbol{y}_n - \boldsymbol{\tau}_0(\boldsymbol{X}_n))/\phi$*, where* $\boldsymbol{\tau}_0(\boldsymbol{X}_n) = (\tau_0(\boldsymbol{x}_1), \ldots, \tau_0(\boldsymbol{x}_n))^\top$*. It comes that*

$$\mathrm{E}[r^{-1}|\boldsymbol{y}_n, \tau_0] = \frac{\nu + d_n}{\nu + n - 2},$$

which is bounded since $d_n = O(n)$.

(3) For slash distribution:

The conditional distribution of r *given* $\boldsymbol{y}_n$ *and* τ_0 *is a truncated gamma distribution* $\mathrm{Gamma}(\nu + n/2, d_n/2)\mathbb{I}_{(0,1)}$*. Then, we get*

$$\begin{aligned}\mathrm{E}[r^{-1}|\boldsymbol{y}_n, \tau_0] &= \frac{d_n}{n + 2\nu - 2} \frac{P_1(\nu + n/2 - 1, d_n/2)}{P_1(\nu + n/2, d_n/2)} \\ &= \frac{d_n}{2} \frac{1}{(\nu + n/2 - 1) - \exp(-d_n/2)/q_n},\end{aligned} \tag{C.17}$$

where

$$\begin{aligned}q_n &= \int_0^1 t^{\nu+n/2-2} e^{-d_n t/2}\,\mathrm{d}t \\ &= (2/d_n)^{\nu+n/2-1} \gamma(\nu + n/2 - 1, d_n/2).\end{aligned} \tag{C.18}$$

Here, $\gamma(a, x) \triangleq \int_0^x t^{a-1} e^{-t}\mathrm{d}t$ *is the incomplete gamma function. Using Theorem 4.1 in Neuman (2013), we find that*

$$q_n \geqslant \frac{1}{\nu + n/2 - 1} \exp\{-\frac{\nu + n/2 - 1}{\nu + n/2} \frac{d_n}{2}\}. \tag{C.19}$$

Combined (C.17) *and* (C.19)*, we have*

$$\begin{aligned}\mathrm{E}[r^{-1}|\boldsymbol{y}_n, \tau_0] &\leqslant \frac{d_n}{(n + 2\nu - 2)(1 - \exp\{-d_n/(n + 2\nu)\})} \\ &\leqslant \frac{d_n}{n + 2\nu - 2} + \frac{n + 2\nu}{n + 2\nu - 2},\end{aligned}$$

which is bounded since $d_n = O(n)$.

(4) For contaminated-normal distribution:

Given $\boldsymbol{y}_n$ and τ_0, r is a discrete random variable with the conditional distribution as $\widetilde{h}(r;\tilde{\nu},\gamma) = \tilde{\nu}\mathbb{I}_{(r=\gamma)} + (1-\tilde{\nu})\mathbb{I}_{(r=1)}$, with $1/\tilde{\nu} = 1 + (1/\nu - 1)\gamma^{-n/2}\exp(-\frac{1-\gamma}{2}d_n)$. Hence, we have $\mathrm{E}[r^{-1}|\boldsymbol{y}_n,\tau_0] = \tilde{\nu}(\gamma^{-1}-1)+1 \leqslant \gamma^{-1}$.

Proof of Theorem 1. Applying Lemma 1 we obtain that

$$\begin{aligned} &\frac{1}{n}\mathrm{E}_{\boldsymbol{X}_n}(\mathrm{D}[p(\boldsymbol{y}_n|\tau_0,\boldsymbol{X}_n),p_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{y}_n|\boldsymbol{X}_n)]) \\ \leqslant &\frac{c}{2n} + \frac{1}{2n}\mathrm{E}_{\boldsymbol{X}_n}(\log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n|) + \frac{b}{2n}(\|\tau_0\|_c^2 + c). \end{aligned} \tag{C.20}$$

Suppose $\|\tau_0\|_c$ is bounded and $\mathrm{E}_{\boldsymbol{X}_n}(\log|\boldsymbol{I}_n + \phi^{-1}\boldsymbol{C}_n|) = o(n)$, then Theorem 1 follows from (C.20).

Remark 2 *The expect regret $\mathrm{E}_{\boldsymbol{X}_n}(\log|\boldsymbol{I}_n+\phi^{-1}\boldsymbol{C}_n|)$ depends both on the covariance function $C(\cdot,\cdot;\boldsymbol{\theta})$ and the distribution $\mathcal{U}(\boldsymbol{x})$, which can be shown as order $o(n)$ for some widely used covariance functions (Seeger et al., 2008).*

References

Cao, C. Z., Lin, J. G., Shi, J. Q., Wang, W., Zhang, X. Y. (2015). Multivariate measurement error models for replicated data under heavy-tailed distributions. Journal of Chemometrics, 29(8), 457-466.

Fang, K. T., Kotz, S., Ng, K. W.(1990). Symmetrical multivariate and related distributions. Chapman and Hall: London.

McLachlan, G. J., Basford, K. E. (1988). Mixture models inference and applications to clustering. Marcel Dekker: New York.

Neuman, E. (2013). Inequalities and bounds for the incomplete gamma function. Results in Mathematics, 63(3), 1209-1214.

Seeger, M. W., Kakade, S. M., Foster, D. P. (2008). Information consistency of nonparametric Gaussian process methods. IEEE transactions on Information Theory, 54(5), 2376-2382.