Supplementary Material for **Bayesian Semiparametric Multivariate Density Deconvolution**

Abhra Sarkar
Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA
abhra.sarkar@duke.edu

Debdeep Pati
Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA
debdeep@stat.fsu.edu

Antik Chakraborty and Bani K. Mallick
Department of Statistics, Texas A&M University, 3143 TAMU, College Station,
TX 77843-3143, USA
antik@stat.tamu.edu and bmallick@stat.tamu.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station,
TX 77843-3143, USA
and School of Mathematical and Physical Sciences, University of Technology Sydney,
Broadway NSW 2007, Australia
carroll@stat.tamu.edu

The Supplementary Material is organized as follows. Section S.1 discusses the choice of hyper-parameters. In Section S.2, we describe a Gibbs sampler for drawing samples from the posterior of the deconvolution model for multivariate independently distributed homoscedastic errors, described in Section 2.2.1 of the main paper. In Section S.3, we detail a two stage estimation procedure for drawing samples from the posterior of the deconvolution model for multivariate conditionally heteroscedastic measurement errors described in Section 2.2.2 of the main paper. Section S.4 provides heuristic justification for the two-stage sampler. In Section S.5, we provide additional detailed discussion of the model for multivariate conditionally heteroscedastic measurement errors described in Section 2.2.2 of the main paper, contrasting it with models for multivariate conditionally varying regression errors (Section S.5.1), its connections with latent factor models (Section S.5.2), its flexibility, limitations, and plausible generalizations (Section S.5.3), and tools for model adequacy checks (Section S.5.4). Section S.6 presents our arguments in favor of finite mixture models, pointing out how their close connections and their subtle differences with possible infinite dimensional alternatives are exploited to achieve significant reduction in computational complexity (Section S.6.2) while retaining the major advantages of infinite dimensional mixture models including model flexibility (Section S.6.4) and automated model selection and model averaging (Section S.6.3). Section S.7 details proofs of the theoretical results presented in Section 5 of the main paper. Section S.8 presents additional figures related to the simulation experiments discussed in Section 6 of the main paper. Section S.9 presents results of additional simulation experiments. Section S.10 discusses potentially far-reaching impact of our work in nutritional epidemiology.

S.1 Choice of Hyper-Parameters

We discuss the choice of hyper-parameters in this section. To avoid unnecessary repetition, in this section and onwards, symbols sans the subscripts $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are sometimes used as generics for similar components and parameters of the models. For example, K is a generic for $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$; $\boldsymbol{\mu}_k$ is a generic for $\boldsymbol{\mu}_{\mathbf{X},k}$ and $\boldsymbol{\mu}_{\boldsymbol{\epsilon},k}$; and so on.

1. **Number of mixture components:** Practical application of our method requires that a decision be made on the number of mixture components $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ in the models for the densities $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, respectively.

Our simulation experiments suggest that when the true densities are finite mixtures of multivariate normals and $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ are assigned values greater than the corresponding true numbers, the MCMC chain often quickly reaches a steady state where the redundant components become empty. See Figures S.6, S.12 and S.13 in the Supplementary Material for illustrations. These observations are similar to that made in the context of ordinary density estimation by Rousseau and Mengersen (2011) who studied the asymptotic behavior of the posterior for overfitted mixture models and showed that when $\alpha/K < L/2$, where L denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components. We set $\alpha_{\mathbf{X}} = \alpha_{\boldsymbol{\epsilon}} = 1$ so that the condition $\alpha/K < L/2$ is satisfied.

Educated guesses about $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ may nevertheless be useful in safeguarding against gross overfitting that would result in a wastage of computation time and resources. The following simple strategies may be employed. Model based cluster analysis techniques as implemented by the mclust package in R (Fraley and Raftery, 2007) may be applied to the starting values of $\mathbf{X}_i$ and the corresponding residuals, obtained by fitting univariate submodels for each component of $\mathbf{X}$, to get some idea about $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$. The chain may be started with larger values of $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ and after a few hundred iterations the redundant empty components may be deleted on the fly.

As shown in Section 5, our methods can approximate a large class of data generating densities, and we found the strategy described above to be very effective in all cases we experimented with. The parameter α now plays the role of a smoothing parameter, smaller values favoring a smaller number of mixture components and thus smoother densities. In simulation experiments involving multivariate t and multivariate Laplace distributions reported in the Supplementary Material, and in some other cases not reported here, the values $\alpha_{\mathbf{X}} = \alpha_{\boldsymbol{\epsilon}} = 1$ worked well.

As we discuss in Section 6, the MIW method becomes highly numerically unstable when the measurement errors are conditionally heteroscedastic and the true covariance matrices are highly sparse. In these cases in particular, the MIW method usually requires much larger sample sizes for the asymptotic results to hold and in finite samples the above mentioned strategy usually overestimates the required number of mixture components. See Figure S.5 in the Supplementary Material for an illustration. Since mixtures based on $(K+1)$ components are at least as flexible as mixtures based on K components, as far as model flexibility is concerned, such overestimation is not an issue. But since this also results in clusters of smaller sizes, the estimates of the component specific covariance matrices become numerically even more unstable, further compounding the stability issues of the MIW model. In contrast, for the numerically more stable MLFA model, for the exact opposite reasons, the asymptotic results are valid for moderate sample sizes and such models are also more robust to overestimation of the number of nonempty clusters.

2. **Number of latent factors:** For the MLFA method, the MCMC algorithm summarized in Section S.2 also requires that the component specific infinite factor models be truncated at some appropriate truncation level. The shrinkage prior again makes the model highly robust to overfitting allowing us to adopt a simple strategy. Since a latent factor characterization leads to a reduction in the number or parameters only when $q_k \leq \lceil (p+1)/2 \rceil$, where $\lceil s \rceil$ denotes the largest integer smaller than or equals to s, we simply set the truncation level at $q_k = q = \max\{2, \lceil (p+1)/2 \rceil\}$ for all the components. We also experimented by setting the truncation level at $q_k = q = p$ for all k with the results remaining practically the same. The shrinkage prior, being continuous in nature, does not set the redundant columns to exact zeroes, but it adaptively shrinks the redundant parameters sufficiently towards zero, thus producing stable and efficient estimates of the densities being modeled.

3. **Other hyper-parameters:** We take an empirical Bayes type approach to assign values to other hyper-parameters. We set $\boldsymbol{\mu}_{\mathbf{X},0} = \overline{\mathbf{X}}^{(0)}$, the overall mean of $\mathbf{X}_{1:n}^{(0)}$, where $\mathbf{X}_{1:n}^{(0)}$ denote the starting values of $\mathbf{X}_{1:n}$ for the MCMC sampler discussed in Section S.2. For the scaled errors we set $\boldsymbol{\mu}_{\boldsymbol{\epsilon},0} = \mathbf{0}$. For the MIW model we take $\nu_0 = (p+2)$, the smallest possible integral value of ν_0 for which the prior mean of $\boldsymbol{\Sigma}_k$ exists. We then take $\boldsymbol{\Sigma}_{\mathbf{X},0}/2 = \boldsymbol{\Psi}_{\mathbf{X},0} = \mathrm{cov}(\overline{\mathbf{X}}_{1:n}^{(0)})$. These choices imply $E(\boldsymbol{\Sigma}_{\mathbf{X},k}) = \boldsymbol{\Psi}_{\mathbf{X},0} = \mathrm{cov}(\overline{\mathbf{X}}^{(0)})$ and, since the variability of each component is expected to be significantly less than the overall variability, ensure noninformativeness. Similarly, for the scaled errors we take $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0}/2 = \boldsymbol{\Psi}_{\boldsymbol{\epsilon},0} = \mathrm{cov}(\boldsymbol{\epsilon}_{1:N}^{(0)})$. For the MLFA model, the hyper-parameters specifying the prior for $\boldsymbol{\Lambda}$ are set at $a_1 = 1, a_h = 2$ for all $h \geq 2$, and $\nu = 1$. Inverse gamma priors with parameters $a_\sigma = 1.1, b_\sigma = 1$ are placed on the elements of $\boldsymbol{\Omega}$. For each k, the variance functions were modeled using quadratic (q=2) B-splines based on $(2 \times 2 + 5 + 1) = 10$ equidistant knot points on $[A_k, B_k] = [\min(\overline{\mathbf{W}}_{k,1:n}) - 0.1 \text{ range}(\overline{\mathbf{W}}_{k,1:n}), \max(\overline{\mathbf{W}}_{k,1:n}) + 0.1 \text{ range}(\overline{\mathbf{W}}_{k,1:n})]$, where $\overline{\mathbf{W}}_{\ell,1:n}$ denotes the subject specific means corresponding to ℓ^{th} component.

S.2 Posterior Computation

Samples from the posterior can be drawn using Gibbs sampling techniques. In what follows $\boldsymbol{\zeta}$ denotes a generic variable that collects the observed proxies $\mathbf{W}_{1:N}$ and all the parameters of a model, including the imputed values of $\mathbf{X}_{1:n}$ and $\boldsymbol{\epsilon}_{1:N}$, that are not explicitly mentioned.

Carefully chosen starting values can facilitate convergence of the sampler. The posterior means of the $X_{i\ell}$'s, obtained by fitting univariate submodels, are used as the starting values for the multivariate sampler. The number of mixture components are initialized at $K_{\mathbf{X}} = (m_{\mathbf{X}} + 2)$, where $m_{\mathbf{X}}$ denotes the optimal number of clusters returned by model based clustering algorithm implemented by the mclust package in R applied to the corresponding initial values $\mathbf{X}_{1:n}^{(0)}$. The component specific mean vectors of the nonempty clusters are set at the mean of $\mathbf{X}_i^{(0)}$ values that belong to that cluster. The component specific mean vectors of the two empty clusters are set at $\overline{\mathbf{X}}^{(0)}$, the overall mean of $\mathbf{X}_{1:n}^{(0)}$. For the MIW model, the initial values of the cluster specific covariance matrices are chosen in a similar fashion. The mixture probabilities for the k^{th} nonempty cluster is set at $\boldsymbol{\pi}_{\mathbf{X},k} = n_k/n$, where n_k denotes the number of $\mathbf{X}_i^{(0)}$ belonging to the k^{th} cluster. The mixture probabilities of the empty clusters are initialized at zero. For the MLFA method, the starting values of all elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}$ are set at zero. The starting values for the elements of $\boldsymbol{\Omega}$ are chosen to equal the variances of the corresponding starting values. The parameters specifying the density of the scaled errors are initialized in a similar manner. The MCMC iterations comprise the following steps. We suppress the subscript $\boldsymbol{\epsilon}$ to keep the notation clean as in the main paper.

1. **Updating the parameters specifying $f_{\mathbf{X}}$:** For the MIW model the parameters specifying the density $f_{\mathbf{X}}$ are updated using the following steps.

$$\begin{aligned}
(\boldsymbol{\pi}|\boldsymbol{\zeta}) &\sim \mathrm{Dir}(\alpha/K + n_1, \alpha/K + n_2, \dots, \alpha/K + n_K),\\
(C_i|\boldsymbol{\zeta}) &\sim \mathrm{Mult}(1, p_{i1}, p_{i2}, \dots, p_{iK}),\\
(\boldsymbol{\mu}_k|\boldsymbol{\zeta}) &\sim \mathrm{MVN}_p(\boldsymbol{\mu}_k^{(n)}, \boldsymbol{\Sigma}_k^{(n)}),\\
(\boldsymbol{\Sigma}_k|\boldsymbol{\zeta}) &\sim \mathrm{IW}_p\{n_k + \nu_0, \textstyle\sum_{i:C_i=k}(\mathbf{X}_i - \boldsymbol{\mu}_k)(\mathbf{X}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} + \boldsymbol{\Psi}_0\},
\end{aligned}$$

where $n_k = \sum_i 1(C_i = k)$, $p_{ik} \propto \pi_k \times \mathrm{MVN}_p(\mathbf{X}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\Sigma}_k^{(n)} = (\boldsymbol{\Sigma}_0^{-1} + n_k\boldsymbol{\Sigma}_k^{-1})^{-1}$ and $\boldsymbol{\mu}_k^{(n)} = \boldsymbol{\Sigma}_k^{(n)}\left\{\boldsymbol{\Sigma}_k^{-1}\sum_{i:C_i=k}\mathbf{X}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right\}$. To update the parameters specifying the covariance matrices in the MLFA model, the sampler cycles through the following steps.

$$\begin{aligned}
(\boldsymbol{\lambda}_{k,j}|\boldsymbol{\zeta}) &\sim \mathrm{MVN}_q\{(\mathbf{D}_{k,j}^{-1} + \sigma_j^{-2}\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{\eta}_k)^{-1}\sigma_j^{-2}\boldsymbol{\eta}_k^{\mathrm{T}}(\mathbf{X}_k^{(j)} - \boldsymbol{\mu}_k^{(j)}), (\mathbf{D}_{k,j}^{-1} + \sigma_j^{-2}\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{\eta}_k)^{-1}\},\\
(\boldsymbol{\eta}_i|C_i = k, \boldsymbol{\zeta}) &\sim \mathrm{MVN}_q\{(\mathrm{I}_q + \boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_k)^{-1}\boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_k), (\mathrm{I}_q + \boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_k)^{-1}\},\\
(\sigma_j^2|\boldsymbol{\zeta}) &\sim \text{Inv-Ga}\left\{a_\sigma + n/2, b_\sigma + (1/2)\textstyle\sum_{i=1}^{n}(X_{ij} - \boldsymbol{\mu}_{C_i,j} - \boldsymbol{\lambda}_{C_i,j}^{\mathrm{T}}\boldsymbol{\eta}_i)^2\right\},\\
(\phi_{k,jh}|\boldsymbol{\zeta}) &\sim \mathrm{Ga}\{(\nu + 1)/2, (\nu + \tau_{k,h}\lambda_{k,jh}^2)/2\},\\
(\delta_{k,h}|\boldsymbol{\zeta}) &\sim \mathrm{Ga}\{a_h + p(q - h + 1)/2, 1 + \textstyle\sum_{\ell=1}^{q}\tau_{k,\ell}^{(h)}\sum_{j=1}^{p}\phi_{k,j\ell}\lambda_{k,j\ell}^2/2\},
\end{aligned}$$

where $D_{k,j}^{-1} = \mathrm{diag}(\phi_{k,j1}\tau_{k,1},\ldots,\phi_{k,jq}\tau_{k,q})$, $\tau_{k,\ell}^{(h)} = \prod_{t=1,t\neq h}^{\ell}\delta_{k,t}$, $\mathbf{X}_k^{(j)} = (X_{i_1 j}, X_{i_2 j},\ldots,X_{i_{n_k} j})^{\mathrm{T}}$, $\boldsymbol{\eta}_k^{n_k\times q} = (\boldsymbol{\eta}_{i_1},\boldsymbol{\eta}_{i_2},\ldots,\boldsymbol{\eta}_{i_{n_k}})^{\mathrm{T}}$, $\{i_1,i_2,\ldots,i_{n_k}\} = \{i : C_i = k\}$.

2. **Updating the parameters specifying $f_{\boldsymbol{\epsilon}}$:** The unconstrained full conditionals of the parameters specifying $f_{\boldsymbol{\epsilon}}$ are very similar. For instance, for the MIW model they are given by

$$\begin{aligned}
(\boldsymbol{\pi}|\boldsymbol{\zeta}) &\sim \mathrm{Dir}(\alpha/K + N_1, \alpha/K + N_2, \ldots, \alpha/K + N_K),\\
(C_{ij}|\boldsymbol{\zeta}) &\sim \mathrm{Mult}(1, p_{ij1}, p_{ij2}, \ldots, p_{ijK}),\\
(\boldsymbol{\mu}_k|\boldsymbol{\zeta}) &\sim \mathrm{MVN}_p(\boldsymbol{\mu}_k^{(N)}, \boldsymbol{\Sigma}_k^{(N)}),\\
(\boldsymbol{\Sigma}_k|\boldsymbol{\zeta}) &\sim \mathrm{IW}_p\{N_k + \nu_0, \textstyle\sum_{ij:C_{ij}=k}(\boldsymbol{\epsilon}_{ij} - \boldsymbol{\mu}_k)(\boldsymbol{\epsilon}_{ij} - \boldsymbol{\mu}_k)^{\mathrm{T}} + \boldsymbol{\Psi}_0\},
\end{aligned}$$

where $N_k = \sum_{i,j} 1(C_{ij} = k)$, $p_{ijk} \propto \pi_k \times \mathrm{MVN}_p(\boldsymbol{\epsilon}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\Sigma}_k^{(N)} = (\boldsymbol{\Sigma}_0^{-1} + N_k\boldsymbol{\Sigma}_k^{-1})^{-1}$ and $\boldsymbol{\mu}_k^{(N)} = \boldsymbol{\Sigma}_k^{(N)}\left\{\boldsymbol{\Sigma}_k^{-1}\sum_{ij:C_{ij}=k}\boldsymbol{\epsilon}_{ij} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right\}$. Samples from the constrained posterior $(\{\boldsymbol{\mu}_k\}_{k=1}^{K}|\sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k = 0, \boldsymbol{\zeta})$ are then obtained from the unconstrained full conditionals $(\boldsymbol{\mu}_k|\boldsymbol{\zeta})$ given above using the simple additional steps described in Section 2.2.2 of the main paper. The steps to update the parameters specifying the covariance matrices in the MLFA model are similarly obtained and are excluded.

3. **Updating the values of X:** When the measurement errors are independent of $\mathbf{X}$, the $\mathbf{X}_i$ have closed form full conditionals given by

$$(\mathbf{X}_i|C_{\mathbf{X},i} = k, C_{\boldsymbol{\epsilon},i1} = k_1, \ldots, C_{\boldsymbol{\epsilon},im_i} = k_{m_i}, \boldsymbol{\zeta}) \sim \mathrm{MVN}_p(\boldsymbol{\mu}_{\mathbf{X}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{X}}^{(n)}),$$

where $\boldsymbol{\Sigma}_{\mathbf{X}}^{(n)} = (\boldsymbol{\Sigma}_{\mathbf{X},k}^{-1} + \sum_{j=1}^{m_i}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}^{-1})^{-1}$ and $\boldsymbol{\mu}_{\mathbf{X}}^{(n)} = \boldsymbol{\Sigma}_{\mathbf{X}}^{(n)}(\boldsymbol{\Sigma}_{\mathbf{X},k}^{-1}\boldsymbol{\mu}_{\mathbf{X},k} + \sum_{j=1}^{m_i}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}^{-1}\mathbf{W}_{ij})$. For conditionally heteroscedastic measurement errors, the full conditionals are given by

$$\begin{aligned}
&(\mathbf{X}_i|C_{\mathbf{X},i} = k, C_{\boldsymbol{\epsilon},i1} = k_1, \ldots, C_{\boldsymbol{\epsilon},im_i} = k_{m_i}, \boldsymbol{\zeta})\\
&\qquad \propto \mathrm{MVN}_p(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k}) \times \textstyle\prod_{j=1}^{m_i}\mathrm{MVN}_p\{\mathbf{W}_{ij}|\mathbf{X}_i + \mathbf{S}(\mathbf{X}_i)\boldsymbol{\mu}_{\boldsymbol{\epsilon},k_j}, \mathbf{S}(\mathbf{X}_i)\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}\mathbf{S}(\mathbf{X}_i)\},
\end{aligned}$$

The full conditionals do not have closed forms. Metropolis-Hastings (MH) steps with multivariate truncated normal proposals are used within the Gibbs sampler.

4. **Updating the parameters specifying s_ℓ:** When the measurement errors are conditionally heteroscedastic, we first estimate the variance functions $s_\ell^2(X_{i\ell})$ by fitting univariate submodels $W_{ij\ell} = X_{i\ell} + s_\ell(X_{i\ell})\epsilon_{ij\ell}$ for each ℓ. The details are provided in Section S.3. The parameters characterizing other components of the full model are then sampled using the Gibbs sampler described above, keeping the estimates of the variance functions fixed.

An alternative class of algorithms integrates out the mixture probabilities $\boldsymbol{\pi}$ and works with the resulting Polya urn scheme (Neal, 2000). We did not consider such algorithms as they render the labels C_i a-priori dependent, requiring the prior conditionals $(C_i|\mathbf{C}_{-i})$ to be recomputed each time any C_i is updated. Importantly, we also need the sampled values of $\boldsymbol{\pi}$ to enforce the zero mean restriction $\sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k = 0$ on the measurement errors.

S.3 Estimation of the Variance Functions

When the measurement errors are conditionally heteroscedastic, we need to update the parameters $\boldsymbol{\xi}_\ell$ that specify the variance functions $s_\ell^2(X_{i\ell})$. These parameters do not have closed form full conditionals. MCMC algorithms, where we tried to integrate MH steps for $\boldsymbol{\xi}_\ell$ with the sampler for the parameters specifying $f_{\boldsymbol{\epsilon}}$, were numerically unstable and failed to converge sufficiently quickly. We need to supply the values of the scaled errors $\epsilon_{ij\ell}$ to step 2 of the algorithm described in Section S.2 and the instability stems from the operation $\boldsymbol{\epsilon}_{ij} = \mathbf{S}(\mathbf{X}_i)^{-1}\mathbf{U}_{ij}$ required to calculate the scaled residuals $\epsilon_{ij\ell}$, as we try to divide $U_{ij\ell}$ by the quantity $s_\ell(X_{i\ell})$, which may be very small for certain values of $X_{i\ell}$, for example, for values of $X_{i\ell}$ near zero for the EATS data application. See Figure 7.

To solve the problem, we adopt a novel two-stage procedure. First, for each k, we estimate the functions $s_\ell^2(X_{i\ell})$ by fitting the univariate submodels $W_{ij\ell} = X_{i\ell} + s_\ell(X_{i\ell})\epsilon_{ij\ell}$. The problem of numerical instability arising out of the operation to determine the values of the scaled errors remains in these univariate subproblems too. But the following lemma from Pelenis (2014), presented here for easy reference, provides us with an escape route by allowing us to avoid this operation in the first place.

Lemma 1. *Let* $\boldsymbol{\theta}_{1:K} = \{(\pi_k, \mu_k, \sigma_k^2)\}_{k=1}^K$ *be such that*

$$f_1(\epsilon|\boldsymbol{\theta}_{1:K}) = \textstyle\sum_{k=1}^K \pi_k \; Normal(\epsilon|\mu_k, \sigma_k^2), \quad with \;\; \sum_{k=1}^K \pi_k = 1, \quad \sum_{k=1}^K \pi_k\mu_k = 0. \tag{S.1}$$

Then there exists a set of parameters $\boldsymbol{\theta}^\star_{1:(K-1)} = \{(\pi_k^\star, p_{k,r}^\star, \mu_{k,r}^\star, \sigma_{k,r}^{\star 2})\}_{r=1,k=1}^{2,K-1}$ *such that*

$$\begin{aligned} f_1(\epsilon|\boldsymbol{\theta}_{1:K}) = f_2(\epsilon|\boldsymbol{\theta}^\star_{1:(K-1)}) &= \textstyle\sum_{k=1}^{K-1} \pi_k^\star \; \sum_{r=1}^2 p_{k,r}^\star Normal(\epsilon|\mu_{k,r}^\star, \sigma_{k,r}^{\star 2}), \\ \textstyle\sum_{k=1}^{K-1} \pi_k^\star = 1, \quad \sum_{r=1}^2 p_{k,r}^\star &= 1, \quad \textstyle\sum_{r=1}^2 p_{k,r}^\star \mu_{k,r}^\star = 0 \;\forall k. \end{aligned} \tag{S.2}$$

Lemma 1 implies that the univariate submodels for the density of the scaled errors given by (S.1) has a reparametrization (S.2) where each component is itself a two-component normal mixture with its mean restricted at zero. The reparametrization (S.2) thus replaces the zero mean restriction on (S.1) by similar restrictions on each of its components. These restrictions also imply that each mixture component in (S.2) can be further reparametrized by only four free parameters. One such parametrization could be in terms of $\widetilde{\boldsymbol{\theta}}_k = (\widetilde{p}_k, \widetilde{\mu}_k, \widetilde{\sigma}_{k,1}^2, \widetilde{\sigma}_{k,2}^2)$, where $(p_{k,1}^\star, \sigma_{k,1}^{\star 2}, \sigma_{k,2}^{\star 2}) = (\widetilde{p}_k, \widetilde{\sigma}_{k,1}^2, \widetilde{\sigma}_{k,2}^2)$ and $\mu_{k,r}^\star = c_{k,r}\widetilde{\mu}_k$, where $c_{k,1} = (1-\widetilde{p}_k)/\{\widetilde{p}_k^2 + (1-\widetilde{p}_k)^2\}^{1/2}$ and $c_{k,2} = -\widetilde{p}_k/\{\widetilde{p}_k^2 + (1-\widetilde{p}_k)^2\}^{1/2}$. Letting p_0 denote the prior assigned to $\widetilde{\boldsymbol{\theta}}_k$, the full conditional of $\widetilde{\boldsymbol{\theta}}_k$ in terms of the conditional likelihood $f_{U|X}$ is proportional to $P_0(\widetilde{\boldsymbol{\theta}}_k)\prod_{ij:C_{\epsilon,ij\ell}=k} f_{U|X}(U_{ij\ell}|X_{i\ell}, \boldsymbol{\xi}_\ell, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\zeta})$. The problem of numerical instability can now be tackled by using MH steps to update not only the parameters $\boldsymbol{\xi}_\ell$ specifying the variance functions but also the parameters $\{\widetilde{\boldsymbol{\theta}}_k\}_k$ characterizing the density f_ϵ using the conditional likelihood $f_{U|X}$ (and not f_ϵ itself), thus escaping the need to separately determine the values of the scaled errors.

The priors and the hyper-parameters for the univariate submodels are chosen following the suggestions of Sarkar, et al. (2014) who used an infinite dimensional extension of this reparametrized finite dimensional submodel. The strategy of exploiting the properties of overfitted mixture models to determine the number of mixture components described in Section S.1 can also be applied to the univariate subproblems. High precision estimates of the variance functions can be obtained using these reparametrized finite dimensional univariate deconvolution models. See Figure 2 and also Figures S.7 and S.16 in the Supplementary Material for illustrations.

A similar reparametrization exists for the multivariate problem too, but the strategy would not be very effective in a multivariate set up as it would require updating the mean vectors and the covariance matrices involved in $f_{\boldsymbol{\epsilon}}$ through MH steps which are not efficient in simultaneous updating of large numbers of parameters. After estimating the parameters characterizing the variance functions from the univariate submodels, we therefore keep these estimates fixed and sample the other parameters using the Gibbs sampler described in Section S.2. Additional details follow.

As discussed in Section 2.2.2 of the main paper, the variance functions s_ℓ^2's can not be uniquely determined without additional identifiability restrictions on the variance of $\epsilon_{ij\ell}$. This, however, does not pose any problem to assess $\text{var}(U_{ij\ell}|X_{i\ell})$ which can be estimated as $\widehat{v}_\ell(X_{i\ell}) = \sum_{m=1}^{M} v_\ell^{(m)}(X_{i\ell})\text{var}^{(m)}(\epsilon_{ij\ell})/M$, where $v_\ell^{(m)}(X_{i\ell})$ and $\text{var}^{(m)}(\epsilon_{ij\ell})$ are estimates of $s_\ell^2(X_{i\ell})$ and $\text{var}(\epsilon_{ij\ell})$ based on the m^{th} sample drawn from the posterior of the ℓ^{th} univariate submodel in the first stage. The final estimate of $\boldsymbol{\xi}_\ell$ is then obtained as $\widehat{\boldsymbol{\xi}}_{\ell,opt} = \arg_{\boldsymbol{\xi}_\ell}\min\sum_{r=1}^{R_\ell}\left\{\widehat{v}_\ell(X_{r\ell}^{\Delta}) - \mathbf{B}_{q,J_\ell,\ell}(X_{r\ell}^{\Delta})\exp(\boldsymbol{\xi}_\ell)\right\}^2$, where $\{X_{r\ell}^{\Delta}\}_{r=1}^{R_\ell}$ is a set of grid points on the support $[A_\ell, B_\ell]$ of the variance functions.

In the second stage, we keep these estimates $\widehat{\boldsymbol{\xi}}_{\ell,opt}$ fixed and sample the other parameters using the Gibbs sampler described in Section S.2. At the m^{th} MCMC iteration of the Gibbs sampler, the scaled errors to be used in step 2 of the algorithm are obtained as $\epsilon_{ij\ell}^{(m)} = (W_{ij\ell} - X_{i\ell}^{(m)})/\widehat{s}_\ell(X_{i\ell}^{(m)})$, where $\widehat{s}_\ell(X_{i\ell}^{(m)}) = \{\mathbf{B}_{q,J_\ell,\ell}(X_{i\ell}^{(m)})\exp(\widehat{\boldsymbol{\xi}}_{\ell,opt})\}^{1/2}$ and $X_{i\ell}^{(m)}$ is sampled value of $X_{i\ell}$ at the m^{th} iteration.

Appropriate scale adjustments are made to make the estimate $\widehat{f}_{\boldsymbol{\epsilon}}$ comparable to the true $f_{\boldsymbol{\epsilon}}$ in simulation experiments. Specifically, $\widehat{f}_{\boldsymbol{\epsilon}} = \sum_{m=1}^{M}\pi_k^{(m)}\text{MVN}(\mathbf{D}\boldsymbol{\mu}_k^{(m)}, \mathbf{D}\boldsymbol{\Sigma}_k^{(m)}\mathbf{D})/M$, where $\mathbf{D} = \text{diag}(\sigma_{true,1}, \dots, \sigma_{true,p})$, $\sigma_{true,\ell}^2$ is the variance of $\epsilon_{ij\ell}$ under the true $f_{\boldsymbol{\epsilon}}$ used to generate them, and $\{\pi_k^{(m)}, \boldsymbol{\mu}_k^{(m)}, \Sigma_k^{(m)}\}_{k=1}^{K}$ are m^{th} sampled values from the posterior of the parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}$ specifying $f_{\boldsymbol{\epsilon}}$.

S.4 The Two-Stage Sampler

Over the last two decades, MCMC techniques have remained at the forefront of Bayesian inference. The literature on the topic is already vast and is still rapidly expanding. While the research on exact MCMC methods is still highly active, owing to numerous practical challenges, approximate computation methods are becoming increasingly popular. For a recent review of traditional exact methods and more recent approximate tools, see Green, et al. (2015). The basic idea of the two-stage sampler described above, while being simple and intuitive, is a novel addition to the growing literature on the topic. We are studying its properties in greater detail in simpler settings in a separate manuscript. Figure S.1 below provides some heuristics.

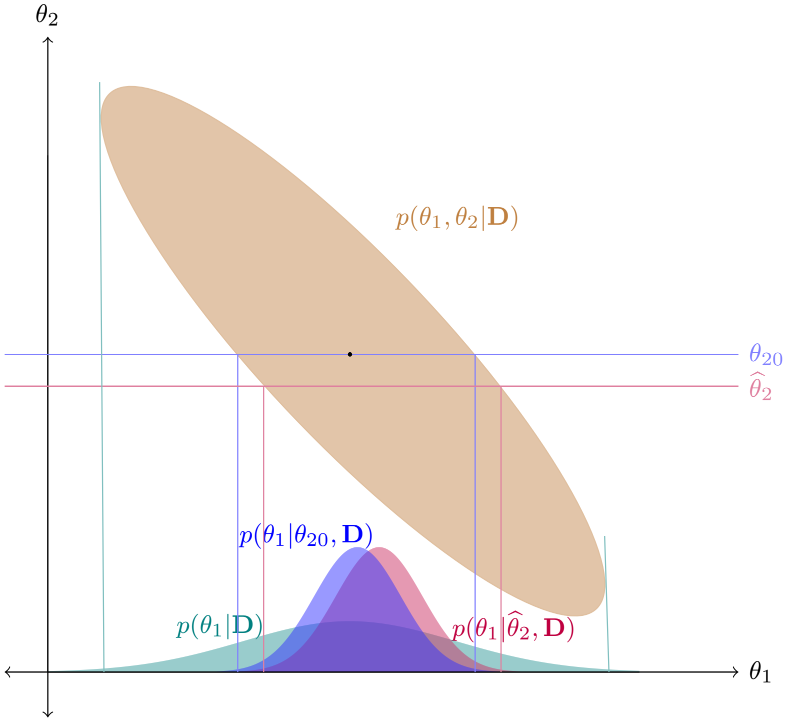

θ_2
$p(\theta_1, \theta_2|\mathbf{D})$
θ_{20}
$\widehat{\theta}_2$
$p(\theta_1|\theta_{20}, \mathbf{D})$
$p(\theta_1|\mathbf{D})$
$p(\theta_1|\widehat{\theta}_2, \mathbf{D})$
θ_1

Figure S.1: Heuristics of the two-stage sampler. The brown elliptical region shows the joint posterior $p(\theta_1, \theta_2|\mathbf{D})$ of two parameters θ_1 and θ_2 given data $\mathbf{D}$. The light blue curve shows $p(\theta_1|\mathbf{D})$, the marginal posterior of θ_1 given data $\mathbf{D}$. The blue curve shows $p(\theta_1|\theta_{20}, \mathbf{D})$, the posterior of θ_1, where θ_{20}, the 'true' value of θ_2, is known. The red curve shows $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$, the pseudo-posterior of θ_1 given $\widehat{\theta}_2$, an estimate of θ_2. $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$ will be close to $p(\theta_1|\theta_{20}, \mathbf{D})$ when $\widehat{\theta}_2$ is close to θ_{20}.

Consider the problem of drawing samples from the posterior $p(\theta_1, \theta_2|\mathbf{D})$ of two parameters θ_1 and θ_2 given data $\mathbf{D}$. The basic MCMC sampler iterates between sampling from (A) $p(\theta_1|\theta_2, \mathbf{D})$ and (B) $p(\theta_2|\theta_1, \mathbf{D})$. If, however, the 'true' value of θ_2 (in a frequentist sense), say θ_{20}, is known, we only require step (A), which becomes $p(\theta_1|\theta_{20}, \mathbf{D})$. And if we substitute θ_2 by a point estimate $\widehat{\theta}_2$, step (A) becomes $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$. While an uncertainty assessment based on $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$ will be overly optimistic compared to that based on the actual marginal posterior $p(\theta_1|\mathbf{D})$, $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$ and $p(\theta_1|\theta_{20}, \mathbf{D})$ will be close when $\widehat{\theta}_2$ is close to θ_{20}, and samples drawn from $p(\theta_1|\widehat{\theta}_2, \mathbf{D})$ may be used for approximate Bayesian inference on θ_1.

The two-stage sampler can also be explained using the following heuristics. Under suitable regularity conditions and considering parametric models (observe that Bayesian nonparametric models are usually large parametric models), the posterior distribution $p(\theta_1, \theta_2|\mathbf{D})$ can be approximated by a Gaussian distribution centered at the true value $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20})$ and variance equal to the inverse of the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$. The justification of this argument is usually tedious and follows from Bernstein von-Mises (BvM) theorems. Refer, for example, to Johnstone (2010), Bontemps (2011), Bickel and Kleijn (2012), Spokoiny (2013) and Castillo and Nickl (2014) for recent literature on BvM theorems in nonparametric Bayesian models and growing parametric Bayesian models. For the sake of convenience, let us assume such results are true for $p(\theta_1, \theta_2|\mathbf{D})$. Hence the marginal posterior distribution $p(\theta_1|\mathbf{D})$ is similar to a Gaussian distribution with mean θ_{10} and variance $[\mathbf{I}(\theta_0)]_{11}^{-1}$, the $(1,1)^{th}$ block of the inverse of $\mathbf{I}(\theta_0)$. Assuming $\widehat{\theta}_2$ to be a consistent estimate of θ_{20}, the conditional posterior distribution in step (A) can be approximated by $p(\theta_1|\theta_{20}, \mathbf{D})$ which in turn is similar to a Gaussian distribution centered at θ_{10} with precision matrix $\mathbf{I}(\theta_{10}|\theta_{20})$, the conditional Fisher information matrix assuming θ_{20} to be known. In classical inference, it is well known that $[\mathbf{I}(\theta_0)]_{11}^{-1} \geq [\mathbf{I}(\theta_{10}|\theta_{20})]^{-1}$ in the sense that the difference is non-negative definite, since knowing θ_{20} results in a higher value of the 'information'. While confidence intervals based on samples drawn by the two-stage algorithm will be optimistic, the draws will be centered around the true value θ_{10} and hence may be used for approximate 'mean' inference on θ_1.

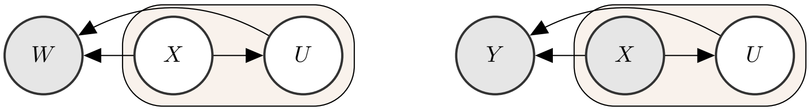
S.5 Comments on the Model for $\mathbf{U}|\mathbf{X}$

As shown in Sarkar, et al. (2014), even in univariate deconvolution settings, due to the non-availability of precise information about X, variations in higher order conditional moments of $(U|X)$ are extremely difficult to capture even in large data sets. Semiparametric approaches that focus separately on the first two moments, namely $E(U|X) = 0$ and $\text{var}(U|X)$, and the shape of $f_{U|X}$, are thus more efficient than possible fully nonparametric approaches even when the truth closely follows the setup of the nonparametric model. See their Section 4.3. This will certainly remain true in the significantly more difficult multivariate deconvolution problem. In building models for $f_{\mathbf{U}|\mathbf{X}}$, we may thus concentrate on the class of models that separates the problem of modeling $\text{cov}(\mathbf{U}|\mathbf{X})$ from that of modeling the shape and other properties of $f_{\mathbf{U}|\mathbf{X}}$. Recent advances in covariance regression models, where the covariance of the multivariate regression errors are allowed to vary flexibly with precisely measured and possibly multivariate predictors, provide us with clues about how this may be achieved. However, as we explain in the following section, there are major differences between conditionally varying multivariate regression errors and conditionally varying multivariate measurement errors. As an implication, covariance regression methods may not be exactly appropriate for modeling conditionally varying covariance matrices $\text{cov}(\mathbf{U}|\mathbf{X})$ in measurement error settings.

S.5.1 Regression Errors vs Measurement Errors

Consider the problem of flexible modeling of conditionally heteroscedastic regression errors where the response and the covariates are both univariate. Consider also the problem of modeling conditionally heteroscedastic measurement errors in a univariate deconvolution set up. From a modeling perspective, Bayesian hierarchical framework allows us to treat these two problems on par by treating both the covariate in the regression problem and the variable of interest in the deconvolution problem simply as conditioning variables. Of course in the regression problem X is precisely measured, whereas in the deconvolution problem X would be latent, but in either case we are required to flexibly model the density of $(U|X)$ subject to $E(U|X) = 0$, where U, depending upon the context, denotes either regression or measurement errors. See Figure S.2. Models for regression errors that allow their variance to vary with the values of the covariate (Pati and Dunson, 2013; Pelenis, 2014) can thus be tried as potential candidates for models for univariate conditionally heteroscedastic measurement errors. Conversely, the models for conditionally heteroscedastic univariate measurement errors (Staudenmayer, et al. 2008; Sarkar, et al. 2014) can also be employed to model univariate conditionally heteroscedastic regression errors.

This is not quite true in a multivariate set up. Interpreting the variables of interest $\mathbf{X}$ broadly as conditioning variables, one can again loosely connect the problem of modeling conditionally heteroscedastic multivariate measurement errors to the problem of covariance



(a) (b)

Figure S.2: (a) Dependency structure in a univariate deconvolution model with latent variable of interest X, associated measurement errors U and replicates W. (b) Dependency structure in a univariate regression model with response Y, associated regression errors U and a univariate observed predictor X. In both panels, the filled rectangular regions focus on the dependency structures between the conditionally varying errors U and the conditioning variable X. The unfilled and the shaded nodes signify latent and observable variables, respectively.

regression (Hoff and Niu, 2012; Fox and Dunson, 2016 etc.), where the goal is to develop models that allow the covariance of multivariate regression errors to vary flexibly with precisely measured and possibly multivariate predictors. In covariance regression problems, the dimension of the regression errors is typically unrelated to the dimension of the predictors. Different components of the regression errors are assumed to be equally influenced by different components of the predictors and hence independent reordering of the components of $\mathbf{X}_i$ will not change the dependency structure. In multivariate deconvolution problems, in contrast, the ℓ^{th} component $U_{ij\ell}$ is the measurement error associated exclusively with $X_{i\ell}$. Here the dimension of $\mathbf{U}_{ij}$ is the same as the dimension of $\mathbf{X}_i$ and any reordering of the components of $\mathbf{X}_i$ would require that the components of $\mathbf{U}_{ij}$ and $\mathbf{W}_{ij}$ be also reordered using the same relabeling scheme. See Figure S.3. While different components of the measurement error vectors $\mathbf{U}_{ij}$ may be correlated, this exclusive association between $U_{ij\ell}$ and $X_{i\ell}$ implies the plausibility that the dependence of $U_{ij\ell}$ on $\mathbf{X}_i$ can be explained primarily through $X_{i\ell}$. Figure 7, for instance, suggests strong conditional heteroscedasticity patterns and it is plausible to assume that the conditional variability in $U_{ij\ell}$ can be explained primarily by $X_{i\ell}$ only. The dependency structure of conditionally varying multivariate measurement errors are, therefore, different from that of conditionally varying multivariate regression errors. Additionally, the aforementioned covariance regression approaches all assume multivariate normality of the regression errors. As is well established in the literature, parametric distributional assumptions on the errors can be particularly restrictive in measurement error problems.

These issues preclude direct application of existing covariance regression approaches to model conditionally heteroscedastic multivariate measurement errors. Models for conditionally varying multivariate measurement errors $(\mathbf{U}|\mathbf{X})$ should highlight their unique features, accommodate distributional flexibility, enforce the mean zero restriction and, to be practically effective, should be computationally stable even in the absence of precise information on the conditioning variable $\mathbf{X}$.

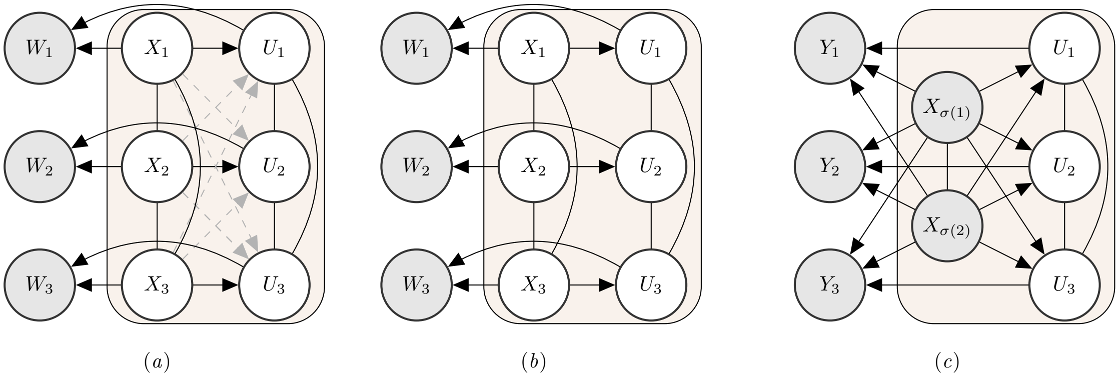


Figure S.3: (a) Dependency structure in a trivariate deconvolution model with latent variable of interest $\mathbf{X} = (X_1, X_2, X_3)^{\mathrm{T}}$, associated measurement errors $\mathbf{U} = (U_1, U_2, U_3)^{\mathrm{T}}$ and replicates $\mathbf{W} = (W_1, W_2, W_3)^{\mathrm{T}}$. The solid black and the dashed gray edges signify strong and weak dependencies, respectively. (b) Dependence relationships in a trivariate deconvolution problem implied by the 'separable' measurement error model $(\mathbf{U}|\mathbf{X}) = \mathbf{S}(\mathbf{X})\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ independent of $\mathbf{X}$ and $\mathbf{S}(\mathbf{X}) = \mathrm{diag}\{s_1(X_1), s_2(X_2), s_3(X_3)\}$. Unlike panel (a), possible weak relationships between U_ℓ and $\{X_m\}_{m\neq\ell}$ are ignored. (c) Dependency structure in a trivariate regression model with response $\mathbf{Y} = (Y_1, Y_2, Y_3)$, associated regression errors $\mathbf{U} = (U_1, U_2, U_3)^{\mathrm{T}}$ and an observed bivariate predictor $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ where $\mathbf{X}_\sigma = (X_{\sigma(1)}, X_{\sigma(2)})^{\mathrm{T}}$ denotes arbitrary reordering of $\mathbf{X}$. In both panels, the filled rectangular regions focus on the dependency structures between the conditionally varying errors $\mathbf{U}$ and the conditioning variable $\mathbf{X}$. The unfilled and the shaded nodes signify latent and observable variables, respectively. The directed and the undirected edges represent one-way and two-way relationships, respectively.

While we reiterate that, for both modeling and computational reasons, the covariance regression methodology of Fox and Dunson (2016) is not be suitable for our purposes, they still provide clues about how the problems of flexible modeling $\mathrm{cov}(\mathbf{U}|\mathbf{X})$ and that of modeling the shape of $f_{\mathbf{U}|\mathbf{X}}$ can be separated. The following section explains.

S.5.2 Latent Factor Models for Different Covariance Classes

Lemma 2 gives a slightly modified version of Lemma 2.1 of Fox and Dunson (2016).

Lemma 2. *Any conditionally varying covariance matrix* $\mathrm{cov}(\mathbf{U}|\mathbf{X}) = \boldsymbol{\Sigma}(\mathbf{X})$ *can be represented as* $\boldsymbol{\Sigma}(\mathbf{X}) = \boldsymbol{\Lambda}(\mathbf{X})\boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{X})$ *for some lower triangular matrix* $\boldsymbol{\Lambda}(\mathbf{X}) = ((\lambda_{\ell,m}(\mathbf{X})))$.

Proof. The proof follows from straightforward application of Cholesky factorization. □

Following Lemma 2, introducing a latent factor $\boldsymbol{\epsilon}$, we can write $(\mathbf{U}|\mathbf{X}, \boldsymbol{\epsilon}) = \boldsymbol{\Lambda}(\mathbf{X})\boldsymbol{\epsilon}$, that is, $(U_\ell|\mathbf{X}, \boldsymbol{\epsilon}) = \sum_{m=1}^{\ell} \lambda_{\ell,m}(\mathbf{X})\epsilon_m$, with $\boldsymbol{\epsilon} \perp \mathbf{X}$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{I}_p$. Completely unrestricted covariance functions can thus be modeled via such latent variable framework by flexibly modeling $\boldsymbol{\Lambda}(\mathbf{X})$. $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$ can be achieved by setting $E(\boldsymbol{\epsilon}) = \mathbf{0}$.

The general nature of the latent factor formulation having been established, we formulate the subsequent results in terms of additional restrictions on such models. Following the discussion in Section S.5.1, we now focus specifically on covariance functions $\mathrm{cov}(\mathbf{U}|\mathbf{X})$ for measurement error problems, where $\mathbf{U}$ and $\mathbf{X}$ are of the same dimension, each component U_ℓ of $\mathbf{U}$ being related to the corresponding component X_ℓ of the conditioning vector $\mathbf{X}$. We consider first the situation when $(U_\ell|\mathbf{X},\boldsymbol{\epsilon})$ depends exclusively on X_ℓ but not on $\{X_m\}_{m\neq\ell}$.

Lemma 3. *Let* $(\mathbf{U}|\mathbf{X},\boldsymbol{\epsilon}) = \boldsymbol{\Lambda}(\mathbf{X})\boldsymbol{\epsilon}$*, where* $\boldsymbol{\Lambda}(\mathbf{X}) = ((\lambda_{\ell,m}(\mathbf{X})))$ *is lower-triangular,* $\boldsymbol{\epsilon} \perp \mathbf{X}$ *and cov*$(\boldsymbol{\epsilon}) = \mathbf{I}_p$*. If* $(U_\ell|\mathbf{X},\boldsymbol{\epsilon}) = (U_\ell|X_\ell,\boldsymbol{\epsilon})$ *for all* ℓ*, then* $\lambda_{\ell,m}(\mathbf{X}) = \lambda_{\ell,m}(X_\ell)$ *for all* ℓ, m*.*

Proof. The proof follows trivially by noting that $(U_\ell|\mathbf{X},\boldsymbol{\epsilon}) = \sum_{m=1}^{\ell}\lambda_{\ell,m}(\mathbf{X})\epsilon_m = (U_\ell|X_\ell,\boldsymbol{\epsilon})$, if and only if, for all $m \le \ell$, $\lambda_{\ell,m}(\mathbf{X})$ is a function of X_ℓ only. □

As an immediate corollary of Lemma 3, the conditional moments $m_\ell^r(\mathbf{X}) = E(U_\ell^r|\mathbf{X})$ are functions of X_ℓ only and the conditional cross-moments $m_{\ell,m}^{r,s}(\mathbf{X}) = E(U_\ell^r U_m^s|\mathbf{X})$ are functions of X_ℓ and X_m only. Modeling variations in the conditional cross-moments is a daunting task in multivariate settings, particularly in the absence of precise information on $\mathbf{X}$. The next result allows the cross-moments $m_{\ell,m}^{r,s}(\mathbf{X})$ to vary with X_ℓ and X_m, but assumes the correlations $\mathrm{corr}(U_\ell,U_m|\mathbf{X})$ to remain constant across $\mathbf{X}$.

Lemma 4. *Let* $(\mathbf{U}|\mathbf{X},\boldsymbol{\epsilon}) = \boldsymbol{\Lambda}(\mathbf{X})\boldsymbol{\epsilon}$*, where* $\boldsymbol{\Lambda}(\mathbf{X}) = ((\lambda_{\ell,m}(\mathbf{X})))$ *is lower-triangular,* $\boldsymbol{\epsilon} \perp \mathbf{X}$ *and cov*$(\boldsymbol{\epsilon}) = \mathbf{I}_p$*. Also, let* $(U_\ell|\mathbf{X},\boldsymbol{\epsilon}) = (U_\ell|X_\ell,\boldsymbol{\epsilon})$ *for all* ℓ*, and corr*$(U_\ell,U_m|\mathbf{X})$ *does not vary with* $\mathbf{X}$ *for all* $\ell \neq m$*. Then,* $\boldsymbol{\Lambda}(\mathbf{X}) = \boldsymbol{\Lambda}_1(\mathbf{X})\mathbf{C}$ *for some diagonal matrix* $\boldsymbol{\Lambda}_1(\mathbf{X}) = \mathrm{diag}\{\lambda_1(X_1),\dots,\lambda_p(X_p)\}$ *and some lower-triangular matrix* $\mathbf{C}$*.*

Proof. From Lemma 3, we have $\lambda_{\ell,m}(\mathbf{X}) = \lambda_{\ell,m}(X_\ell)$ for all ℓ, m, and $\mathrm{corr}(U_\ell,U_m|\mathbf{X})$ varies with X_ℓ and X_m only. Under the additional assumption of Lemma 4, we first prove that $\lambda_{\ell,m}(X_\ell) = c_{\ell,m}\lambda_{\ell,\ell}(X_\ell)$ for some constant $c_{\ell,m}$ for all $m < \ell$ and all $\ell = 2,\dots,p$. Without loss of generality, we assume that $\mathrm{corr}(U_\ell,U_m|\mathbf{X}) = r_{\ell,m} \neq 0$ for all $\ell \neq m$. We have

$$\mathrm{corr}(U_1,U_2|\mathbf{X}) = \frac{\lambda_{2,1}(X_2)}{\{\lambda_{2,1}^2(X_2)+\lambda_{2,2}^2(X_2)\}^{1/2}} = r_{1,2} \;\Rightarrow\; \lambda_{2,2}^2(X_2) = \frac{(1-r_{1,2}^2)}{r_{1,2}^2}\lambda_{2,1}^2(X_2). \quad \text{(S.3)}$$

So the proposition holds true for $\ell = 2$. Next, assume that it holds for $\ell = 2,\dots,h-1$ for some $h > 2$. Also, from (S.3), $\mathrm{var}(U_2|\mathbf{X}) = \sum_{m=1}^{2}\lambda_{2,m}^2(X_2) = \lambda_{2,1}^2(X_2)/r_{1,2}^2$. This is, in fact, more generally true for all ℓ. For instance, for $\ell = h$,

$$\mathrm{corr}(U_1,U_h|\mathbf{X}) = \frac{\lambda_{h,1}(X_h)}{\{\sum_{m=1}^{h}\lambda_{h,m}^2(X_h)\}^{1/2}} = r_{1,h} \quad\Rightarrow\quad \sum_{m=2}^{h}\lambda_{h,m}^2(X_h) = \frac{(1-r_{1,h}^2)}{r_{1,h}^2}\lambda_{h,1}^2(X_h)$$

$$\Rightarrow\quad \mathrm{var}(U_h|\mathbf{X}) = \textstyle\sum_{m=1}^{h}\lambda_{h,m}^2(X_h) = \lambda_{h,1}^2(X_h)/r_{1,h}^2. \quad \text{(S.4)}$$

$$\text{Then,}\quad \mathrm{corr}(U_2,U_h|\mathbf{X}) = \frac{\lambda_{2,1}(X_2)\lambda_{h,1}(X_h)+\lambda_{2,2}(X_2)\lambda_{h,2}(X_h)}{\{\sum_{m=1}^{2}\lambda_{2,m}^2(X_2)\}^{1/2}\{\sum_{m=1}^{h}\lambda_{h,m}^2(X_h)\}^{1/2}} = r_{2,h}$$

$$\Rightarrow \quad \frac{\lambda_{2,2}(X_2)\{c_{2,1}\lambda_{h,1}(X_h)+\lambda_{h,2}(X_h)\}}{|c_{2,1}\lambda_{2,2}(X_2)|\,|\lambda_{h,1}(X_h)|} = \frac{r_{2,h}}{|r_{1,2}r_{1,h}|}.$$

$$\Rightarrow \quad \lambda_{h,2}(X_h) = \widetilde{c}_{h,2}\lambda_{h,1}(X_h) \text{ for some constant } \widetilde{c}_{h,2}. \tag{S.5}$$

$$\text{Next,} \quad \mathrm{corr}(U_3,U_h|\mathbf{X}) = \frac{\sum_{m=1}^{3}\lambda_{3,m}(X_3)\lambda_{h,m}(X_h)}{\{\sum_{m=1}^{3}\lambda_{3,m}^2(X_3)\}^{1/2}\{\sum_{m=1}^{h}\lambda_{h,m}^2(X_h)\}^{1/2}} = r_{3,h}$$

$$\Rightarrow \quad \frac{\lambda_{3,3}(X_3)\{c_{3,1}\lambda_{h,1}(X_h)+c_{3,2}\widetilde{c}_{h,2}\lambda_{h,1}(X_h)+\lambda_{h,3}(X_h)\}}{|c_{3,1}\lambda_{3,3}(X_3)|\,|\lambda_{h,1}(X_h)|} = \frac{r_{3,h}}{|r_{1,3}r_{1,h}|}$$

$$\Rightarrow \quad \lambda_{h,3}(X_h) = \widetilde{c}_{h,3}\lambda_{h,1}(X_h) \text{ for some constant } \widetilde{c}_{h,3}. \tag{S.6}$$

$$\text{Finally,} \quad \mathrm{corr}(U_{h-1},U_h|\mathbf{X}) = \frac{\sum_{m=1}^{h-1}\lambda_{h-1,m}(X_{h-1})\lambda_{h,m}(X_h)}{\{\sum_{m=1}^{h-1}\lambda_{h-1,m}^2(X_{h-1})\}^{1/2}\{\sum_{m=1}^{h}\lambda_{h,m}^2(X_h)\}^{1/2}} = r_{h-1,h}$$

$$\Rightarrow \quad \frac{\lambda_{h-1,h-1}(X_{h-1})\{c_{h-1,1}\lambda_{h,1}(X_h)+c_{h-1,2}\widetilde{c}_{h,2}\lambda_{h,1}(X_h)+\cdots+\lambda_{h,h}(X_h)\}}{|c_{h-1,1}\lambda_{h-1,1}(X_{h-1})|\,|\lambda_{h,1}(X_h)|} = \frac{r_{h-1,h}}{|r_{1,h-1}r_{1,h}|}$$

$$\Rightarrow \quad \lambda_{h,h-1}(X_h) = \widetilde{c}_{h,h-1}\lambda_{h,1}(X_h) \text{ for some constant } \widetilde{c}_{h,h-1}. \tag{S.7}$$

Combining (S.5), (S.6), (S.7) etc. with (S.4), the proposition follows by principles of mathematical induction. This implies $\mathbf{\Lambda}(\mathbf{X}) = \mathbf{\Lambda}_1(\mathbf{X})\mathbf{C}$ where $\mathbf{\Lambda}_1(\mathbf{X}) = \mathrm{diag}\{\lambda_1(X_1),\dots,\lambda_p(X_p)\}$ with $\lambda_\ell(X_\ell) = \lambda_{\ell,\ell}(X_\ell)$ for all ℓ and $\mathbf{C} = ((c_{\ell,m}))$ is a lower triangular matrix with $c_{\ell,\ell} = 1$ for all ℓ. □

Under the conditions of Lemma 4, we thus have $\mathrm{cov}(\mathbf{U}|\mathbf{X}) = \mathbf{\Sigma}(\mathbf{X}) = \mathbf{\Lambda}_1(\mathbf{X})\mathbf{\Sigma}_1\mathbf{\Lambda}_1^{\mathrm{T}}(\mathbf{X})$ with $\mathbf{\Sigma}_1 = \mathbf{C}\mathbf{C}^{\mathrm{T}}$. Introducing a latent factor $\boldsymbol{\epsilon}$, we can now write $(\mathbf{U}|\mathbf{X},\boldsymbol{\epsilon}) = \mathbf{\Lambda}_1(\mathbf{X})\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \perp \mathbf{X}$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{\Sigma}_1$. Due to the diagonal nature of $\mathbf{\Lambda}_1(\mathbf{X})$, each component ϵ_ℓ of $\boldsymbol{\epsilon}$ is exclusively associated with the corresponding component U_ℓ of $\mathbf{U}$ and may be treated as a scaled version of U_ℓ. Starting with a general latent factor model framework, with two additional restrictions that are particularly relevant in multivariate measurement error settings, we have now arrived at model (16). The problems of modeling $\mathrm{cov}(\mathbf{U}|\mathbf{X})$ and the shape of $f_{\mathbf{U}|\mathbf{X}}$ can now be achieved by separately modeling $\mathbf{\Lambda}_1(\mathbf{X})$ and $f_{\boldsymbol{\epsilon}}$. And $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$ can be achieved by enforcing $E(\boldsymbol{\epsilon}) = \mathbf{0}$.

S.5.3 Models for U|X and cov(U|X)

In this section, we first revisit the models for conditionally varying measurement errors developed in Section 2.2 of the main paper. A few plausible alternatives and generalizations, the implied covariance structures, their strengths, limitations and connections with the adopted model are also discussed.

The model (16) for conditionally varying measurement errors developed in Section 2.2 of the main paper assumes $(\mathbf{U}_{ij}|\mathbf{X}_i) = \mathbf{S}(\mathbf{X}_i)\boldsymbol{\epsilon}_{ij\ell}$ where $\mathbf{S}(\mathbf{X}_i) = \mathrm{diag}\{s_1(X_{i1}),\dots,s_p(X_{ip})\}$ and $\boldsymbol{\epsilon}_{ij\ell}$ are distributed independently of $\mathbf{X}$ with $E(\boldsymbol{\epsilon}_{ij}) = \mathbf{0}$. This 'separability' of $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_{ij}$ allows us to incorporate distributional flexibility and enforce the mean zero restriction using the techniques developed for independent errors in Section 2.2.1 in the main paper. The diagonal structure of $\mathbf{S}$ highlights the exclusive associations between $U_{ij\ell}$ and $X_{i\ell}$ but ignores weak dependencies of $U_{ij\ell}$ on $\{X_{im}\}_{m\neq\ell}$. The general of shape of $f_{\mathbf{U}|\mathbf{X}}$ as well correlations between different components of $\mathbf{U}_{ij}$ are inherited from $f_{\boldsymbol{\epsilon}}$. The associated dependency structure is summarized in Figure S.3(b). The novel two-stage procedure described in Sections S.2 and S.3 produces efficient and numerically stable posterior estimates.

As discussed in Section 2.2.3, the model also arises naturally in multivariate multiplicative measurement error settings $\mathbf{W}_{ij} = \mathbf{X}_i \circ \widetilde{\mathbf{U}}_{ij}$ where the errors $\widetilde{\mathbf{U}}_{ij}$ are distributed independently of $\mathbf{X}_i$ with $E(\widetilde{\mathbf{U}}_{ij}) = \mathbf{1}$. The model can be reformulated as $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, where $\mathbf{U}_{ij} = \mathbf{S}(\mathbf{X}_i)\boldsymbol{\epsilon}_{ij}$, $\mathbf{S}(\mathbf{X}_i) = \mathrm{diag}\{X_{i1},\dots,X_{ip}\}$ and $\boldsymbol{\epsilon}_{ij} = (\widetilde{\mathbf{U}}_{ij} - 1)$ with $E(\boldsymbol{\epsilon}_{ij}) = \mathbf{0}$. It thus conforms to the conditionally varying additive measurement error model (16) described above.

These results and the ones provided in Section S.5.2 establish the fairly general nature of model (16) and are also informative about cases outside its support. A few such cases that are particularly relevant to measurement error problems and form part of our research aspirations but are not pursued in detail in this article are briefly discussed below.

As informed by Lemma 3, another class that implies $\mathrm{var}(U_{ij\ell}|\mathbf{X}_i) = s_\ell^2(X_{i\ell})$ and allows $\mathrm{corr}(U_{ij\ell}, U_{ijm}|\mathbf{X}_i)$ to vary with $X_{i\ell}$ and X_{im} is obtained by letting $\mathbf{U}_{ij} = \boldsymbol{\Lambda}(\mathbf{X}_i)\boldsymbol{\epsilon}_{ij}$ with $\boldsymbol{\Lambda}(\mathbf{X}_i) = ((\lambda_{\ell,m}(X_{i\ell})))_{\ell=1,m=1}^{p,p}$. The model highlights the exclusive associations between $U_{ij\ell}$ and $X_{i\ell}$ - $\mathrm{var}(U_{ij\ell}|\mathbf{X}_i)$ depends on $X_{i\ell}$ and $\mathrm{cov}(U_{ij\ell}, U_{ijm}|\mathbf{X}_i)$ depends on $X_{i\ell}$ and X_{im}. Modeling variations in conditional cross-moments is a daunting task in multivariate settings, more so in the absence of precise information about $\mathbf{X}_i$. Towards a more parsimonious representation, the off-diagonal elements $\{\lambda_{\ell,m}(X_{i\ell})\}_{\ell\neq m}$ may be shrunk towards zero, resulting in a model that associates each $U_{ij\ell}$ with its own latent factor component $\epsilon_{ij\ell}$. That is, $\boldsymbol{\Lambda}(\mathbf{X}_i)$ should be shrunk towards $\boldsymbol{\Lambda}_0(\mathbf{X}_i) = \mathrm{diag}\{\lambda_{1,1}(X_{i1}),\dots,\lambda_{p,p}(X_{ip})\}$. This limiting case still allows $\mathrm{var}(U_{ij\ell}|\mathbf{X}_i)$ to vary flexibly with $X_{i\ell}$, and $\mathrm{cov}(U_{ij\ell}, U_{ijm}|\mathbf{X})$ to vary with $X_{i\ell}$ and X_{im}, but assumes the correlations $\mathrm{corr}(U_{ij\ell}, U_{ijm}|\mathbf{X}_i)$ to not vary with $\mathbf{X}_i$.

Another flexible class of models for $(\mathbf{U}_{ij}|\mathbf{X}_i)$ that conforms to the dependency structure depicted in Figure S.3(a) is obtained by letting $\mathbf{U}_{ij} = \boldsymbol{\Lambda}(\mathbf{X}_i)\boldsymbol{\epsilon}_{ij}$ with $\boldsymbol{\Lambda}(\mathbf{X}_i) = ((\lambda_{\ell,m}(X_{im})))_{\ell=1,m=1}^{p,p}$. The implied covariance structure is given by $\mathrm{cov}(\mathbf{U}_{ij}|\mathbf{X}_i) = \boldsymbol{\Sigma}(\mathbf{X}_i) = \boldsymbol{\Lambda}(\mathbf{X}_i)\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{X}_i)$. Specifically, we have $(U_{ij\ell}|\mathbf{X}_i) = \sum_m \lambda_{\ell,m}(X_{im})\epsilon_{ijm}$ with

$$\begin{aligned}
\mathrm{cov}(U_{ij\ell_1}, U_{ij\ell_2}|\mathbf{X}_i) &= \textstyle\sum_{m_1,m_2} \lambda_{\ell_1,m_1}(X_{im_1})\lambda_{\ell_2,m_2}(X_{im_2})\sigma_{m_1,m_2} \\
&= \lambda_{\ell_1,\ell_1}(X_{i\ell_1})\lambda_{\ell_2,\ell_2}(X_{i\ell_2})\sigma_{\ell_1,\ell_2} + \textstyle\sum_{m_1\neq\ell_1,m_2\neq\ell_2} \lambda_{\ell_1,m_1}(X_{im_1})\lambda_{\ell_2,m_2}(X_{im_2})\sigma_{m_1,m_2} \\
\text{and} \quad \mathrm{var}(U_{ij\ell}|\mathbf{X}_i) &= \lambda_{\ell,\ell}^2(X_{i\ell})\sigma_{\ell,\ell} + \textstyle\sum_{m_1\neq\ell,m_2\neq\ell} \lambda_{\ell,m_1}(X_{im_1})\lambda_{\ell,m_2}(X_{im_2})\sigma_{m_1,m_2}.
\end{aligned}$$

Ideally, to highlight the exclusive strong association between $U_{ij\ell}$ and $X_{i\ell}$, the diagonal elements of $\boldsymbol{\Lambda}(\mathbf{X}_i)$, namely $\lambda_{\ell,\ell}(X_{i\ell})$, should dominate and the remaining off-diagonal elements $\{\lambda_{\ell,m}(X_{im})\}_{\ell\neq m}$ may be shrunk towards zero. That is, $\boldsymbol{\Lambda}(\mathbf{X}_i)$ should be shrunk towards $\boldsymbol{\Lambda}_0(\mathbf{X}_i) = \mathrm{diag}\{\lambda_{1,1}(X_{i1}),\dots,\lambda_{p,p}(X_{ip})\}$.

Since measurement error problems are well known to be inherently computationally unstable, it is not clear whether any practical gain in efficiency can be achieved by modeling large number of off-diagonal functions in $\mathbf{\Lambda}(\mathbf{X}_i)$ at the expense of significantly increased model complexity. Model (16) considered in this article instead focuses on the special limiting cases with $\mathbf{S}(\mathbf{X}_i) = \mathbf{\Lambda}_0(\mathbf{X}_i)$.

Another extension results from mixtures of multiplicative and independent additive errors. In univariate settings, such models were considered in Rocke a Durbin (2001) for studying gene expression levels measured by DNA slides. In multivariate settings, we have $\mathbf{U}_{ij} = \mathbf{X}_i \circ \boldsymbol{\epsilon}_{ij}^{(1)} + \boldsymbol{\epsilon}_{ij}^{(2)}$, where $\boldsymbol{\epsilon}_{ij}^{(k)}$, $k = 1, 2$ are distributed independently of $\mathbf{X}_i$. With $\mathrm{cov}(\boldsymbol{\epsilon}_{ij}^{(k)}) = \mathbf{\Sigma}_k = ((\sigma_{\ell,m}^{(k)}))_{m=1,\ell=1}^{p,p}$ for $k = 1, 2$, the implied covariance structure is given by $\mathrm{cov}(\mathbf{U}_{ij}|\mathbf{X}_i) = \mathbf{S}(\mathbf{X}_i)\mathbf{\Sigma}_1\mathbf{S}(\mathbf{X}_i) + \mathbf{\Sigma}_2$, where $\mathbf{S}(\mathbf{X}_i) = \mathrm{diag}\{X_{i1}, \ldots, X_{ip}\}$, as above. The model conforms to the dependency structure of Figure S.3(b) but can not be strictly written as model (16). However, as can be seen from Figure 7, in our motivating nutritional epidemiology application, smaller average consumptions naturally result in more precise 24 hour recalls, the variability approaching 0 as the true consumption approaches 0. Under the assumption of continuity, $\lim_{\mathbf{X}\to\mathbf{0}} \mathbf{\Sigma}(\mathbf{X}) \to \mathbf{0}^{p\times p}$ implies $\mathbf{\Sigma}_2 = \mathbf{0}^{p\times p}$, resulting in model (16).

S.5.4 Model Adequacy Checks

In Figure 7 in the main paper, we showed the plots of subject specific means $\overline{W}_{i\ell}$ of the replicates vs the corresponding subject-specific variances $S_{W,i\ell}^2$ for each of the four dietary components included in our analysis in Section 7. These plots suggest very strong conditional heteroscedasticity patterns in the measurement errors. If we consider the plots of subject specific means $\overline{W}_{i\ell}$ vs subject specific variances $S_{W,im}^2$ for all possible pairs (ℓ, m), we will see similar monotone increasing patterns not just for the pairs with $\ell = m$, but in pairs with $\ell \neq m$ too. This can be explained by the high correlation between different components of $\mathbf{X}_i$, see Figure 8, and does not necessarily imply that the conditional variability in $U_{ij\ell}$ depends on other components of $\mathbf{X}_i$, not just $X_{i\ell}$. As discussed in the previous subsections, since the ℓ^{th} component $U_{ij\ell}$ is the measurement error associated exclusively with $X_{i\ell}$, it is plausible to assume that the conditional variability of $U_{ij\ell}$ can be modeled mostly as a function of $X_{i\ell}$ only.

We present here some diagnostic plots to further validate the practical adequacy of this structural assumption. Figure S.4 shows the plots of $\widehat{X}_{i\ell}$ vs subject specific variances $\widehat{S}_{\epsilon,im}^2$ of $\widehat{\epsilon}_{ijm}$, where $\widehat{X}_{i\ell}$ represent the posterior means of $X_{i\ell}$ values and $\widehat{\epsilon}_{ijm} = (W_{ijm} - \widehat{X}_{im})/\widehat{s}_m(\widehat{X}_{im})$ represent the corresponding scaled measurement error residuals produced by the univariate submodels for the EATS data set analyzed in Section 7 of the main paper. The figure indicates constant variance of the scaled measurement error residuals $\widehat{\epsilon}_{ij\ell}$ over the entire range of X_{im} values for all (ℓ, m) pairs. Nonparametric Eubank-Hart tests of no covariate effect (Eubank and Hart, 1992) applied to $(\widehat{X}_{i\ell}, \widehat{S}^2_{\epsilon,im})$ for all (ℓ, m) pairs (treating $\widehat{X}_{i\ell}$ as the covariate and $\widehat{S}^2_{\epsilon,im}$ as the response) produced a minimum Benjamini-Hochberg adjusted p-value of 0.096, suggesting that there is no residual heteroscedasticity left in $U_{ij\ell}$ after accounting for the variability in $U_{ij\ell}$ that can be sufficiently explained through $X_{i\ell}$ only. See Table S.1. It may thus be concluded that for the EATS data application model (16) developed in Section 2.2.2 of the main paper that implies $\mathrm{var}(U_{ij\ell}|\mathbf{X}_i) = s^2_\ell(X_{i\ell})\mathrm{var}(\epsilon_{ij\ell})$ suffices to explain the conditional variability in the measurement errors.

Model (16) also assumed that only the conditional variability of $\mathbf{U}_{ij}$ depends on $\mathbf{X}_i$, and derived other features of $\mathbf{U}_{ij}$ like skewness, multimodality, heavy-tails etc. from the scaled errors $\boldsymbol{\epsilon}_{ij}$. As shown in Sarkar, et al. (2014), even in the much simpler univariate set up, in the absence of precise information on $X_{i\ell}$, variations in other features of $U_{ij\ell}$ for varying values of $X_{i\ell}$, if any, are extremely difficult to detect. More importantly, semiparametric methods that make the multiplicative structural assumption $(U_{ij\ell}|X_{i\ell}) = s_\ell(X_{i\ell})\epsilon_{ij\ell}$ are highly robust to departures from this assumption and significantly outperform possible nonparametric alternatives that allow all order moments of $U_{ij\ell}$ to vary flexibly with $X_{i\ell}$, not just the conditional variance, even in scenarios where the true data generating process closely conforms to these nonparametric alternatives.

	Panel	p-values	BFN	BH	BY
1	1,1	0.991	1.000	0.991	1.000
2	1,2	0.764	1.000	0.873	1.000
3	1,3	0.251	1.000	0.446	1.000
4	1,4	0.129	1.000	0.446	1.000
5	2,1	0.598	1.000	0.736	1.000
6	2,2	0.266	1.000	0.446	1.000
7	2,3	0.037	0.592	0.197	0.667
8	2,4	0.990	1.000	0.991	1.000
9	3,1	0.224	1.000	0.446	1.000
10	3,2	0.012	0.192	0.096	0.325
11	3,3	**0.011**	**0.176**	**0.096**	**0.325**
12	3,4	0.497	1.000	0.692	1.000
13	4,1	0.519	1.000	0.692	1.000
14	4,2	0.163	1.000	0.446	1.000
15	4,3	0.279	1.000	0.446	1.000
16	4,4	0.244	1.000	0.446	1.000

Table S.1: The original and adjusted p-values (BFN=Bonferroni, BH=Benjamini-Hochberg, BY=Benjamini-Yekutli) returned by nonparametric Eubank-Hart tests of no covariate effect applied to $(\widehat{X}_{i\ell}, \widehat{S}^2_{\epsilon,im})$ for all (ℓ, m) pairs treating $\widehat{X}_{i\ell}$ as the covariate and $\widehat{S}^2_{\epsilon,im}$ as the response. The minimum values corresponding to panel $(3, 3)$ are highlighted. See Section S.5.4 and Figure S.4 in the Supplementary Material for additional details.

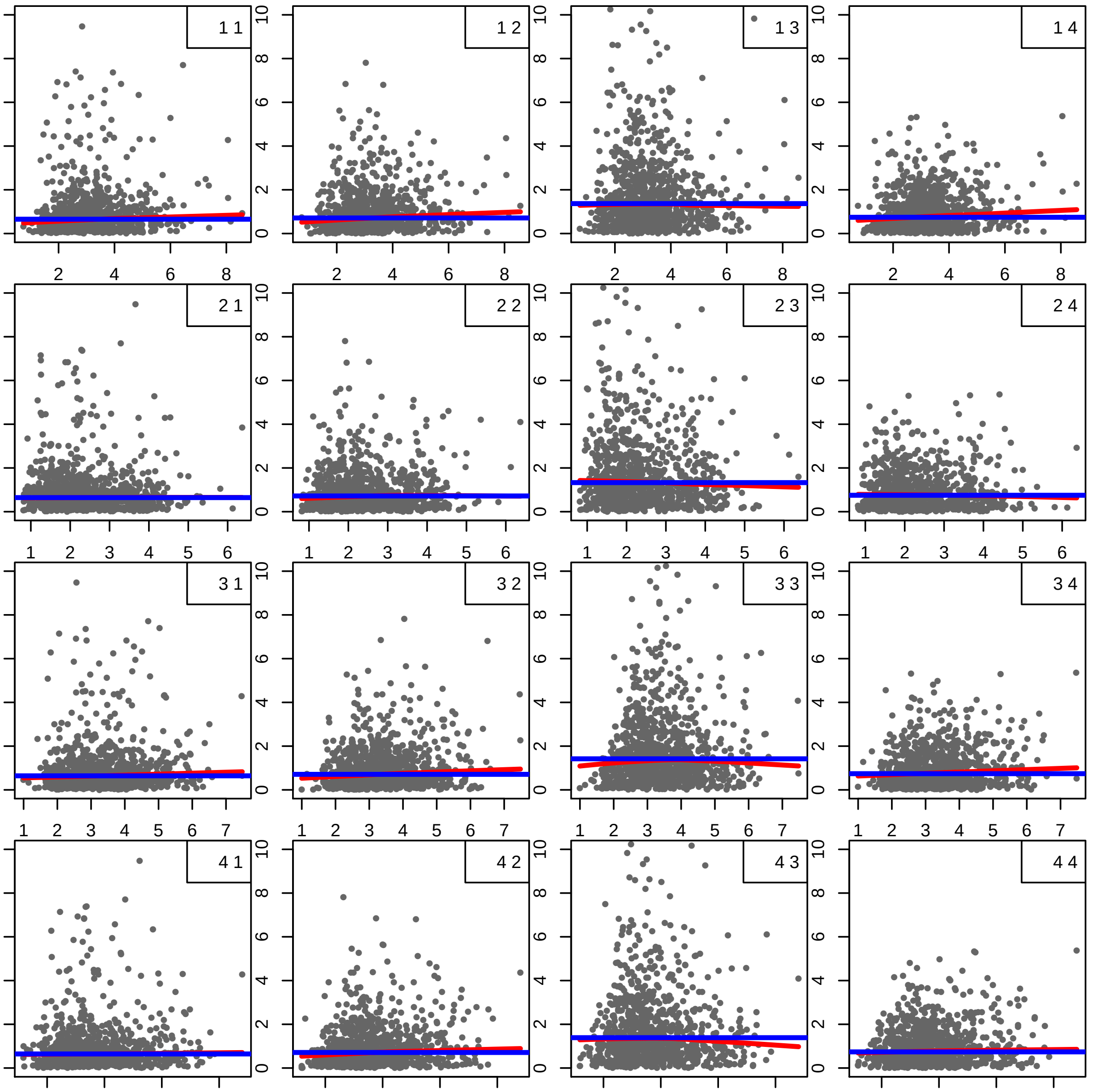


Figure S.4: Panel (ℓ, m) shows the plot of estimates $\widehat{X}_{i\ell}$ of $X_{i\ell}$ vs subject specific variances $\widehat{S}^2_{\epsilon,im}$ of scaled measurement error residuals $\widehat{\epsilon}_{ijm}$, produced by univariate deconvolution methods. See Section S.5.4 of the Supplementary Material for additional details. The darker horizontal lines in each panel represent the upper 10% trimmed mean of the subject specific variances $\widehat{S}^2_{\epsilon,i\ell}$. The lighter solid lines in each panel represent nonparametric lowess fits.

S.6 Finite vs Infinite Mixture Models

In this article, we modeled the $f_{\mathbf{X}}$ and the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$ using mixtures of fixed finite number of multivariate normal kernels. Alternative approaches that escape the need to prespecify the number of mixture components include models with potentially infinite number of mixture components, models induced by Dirichlet processes (Ferguson, 1973; Escobar and West, 1995) being perhaps the most popular among such techniques. Apart from flexibility, one major advantage of such techniques comes from the ability of associated MCMC machinery to perform model selection and model averaging implicitly and semiautomatically. Model averaging is achieved by allowing the number of mixture components to vary from one MCMC iteration to the other. The number of mixture components that is visited the maximum number of times by the sampler then provides a maximum a-posteriori (MAP) estimate of the number of mixture components required to approximate the target density. However, in complicated multivariate set up like ours, MCMC algorithms for such infinite dimensional models become computationally highly intensive. Mixtures based on fixed finite number of components, on the other hand, can greatly reduce computational complexity. Recent studies of asymptotic properties of the posterior of overfitted mixture models (Rousseau and Mengersen, 2011) suggest that mixture models with sufficiently large number of components can perform automatic model selection and model averaging just like infinite dimensional models. Additionally, as the proofs of the results in Section 5 imply, the use of mixture models with fixed finite number of components does not necessarily imply a compromise on the issue of flexibility. The approaches adopted in this article try to take the best from both worlds. Computational burden is reduced by keeping the number of mixture components fixed at some finite values. At the same time, simultaneous semiautomatic model selection and model averaging is achieved by exploiting properties of overfitted mixture models. We elaborate our arguments below, pointing out the close connections and the subtle differences our adopted finite dimensional models have with the aforementioned infinite dimensional alternatives.

S.6.1 Infinite Mixture Models as Limits of Finite Mixture Models

Let $G_K = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$ with $(\pi_1, \dots, \pi_K) \sim \mathrm{Dir}(\alpha/K, \dots, \alpha/K)$ and $\theta_k \sim H$. Also, let $G_\infty \sim \mathrm{DP}(\alpha, H)$, a Dirichlet process with concentration parameter α and base measure H. Then, G_∞ can be represented as $G_\infty = \sum_{k=1}^{\infty} \widetilde{\pi}_k \delta_{\theta_k}$ with $\widetilde{\pi}_k = V_k \prod_{\ell=1}^{k-1}(1 - V_\ell), V_\ell \sim \mathrm{Beta}(1, \alpha)$ and $\theta_k \sim H$ (Sethuraman, 1994). As $K \to \infty$, $\int g(\theta) dG_K(\theta) \overset{d}{\to} \int g(\theta) dG_\infty(\theta)$ for any measurable function g integrable with respect to H (Ishwaran and Zarepour, 2000, 2002).

The finite mixtures of multivariate normal kernels with symmetric Dirichlet priors that we used in this article to model both $f_{\mathbf{X}}$ and the density of the scaled measurement errors

$f_{\boldsymbol{\epsilon}}$ have close connections with infinite dimensional Dirichlet process based mixture models. Specifically, taking $g(\theta) = \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and appealing to the above result, we have $f_{\mathbf{X}} = \sum_{k=1}^{K_{\mathbf{X}}} \pi_{\mathbf{X},k} \mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k}) \overset{d}{\to} \sum_{k=1}^{\infty} \widetilde{\pi}_{\mathbf{X},k} \mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k})$ as $K_{\mathbf{X}} \to \infty$. Our proposed mechanism to enforce the mean zero restriction on $f_{\boldsymbol{\epsilon}}$ specifically requires a finite dimensional symmetric prior on the mixture probabilities and therefore does not admit a straightforward infinite dimensional extension. But in the limit, as $K_{\boldsymbol{\epsilon}} \to \infty$, a reformulation of the model results in a complicated multivariate version of the infinite dimensional model of Sarkar, et al. (2014) (See Lemma 1 in Section S.3).

S.6.2 Computational Complexity

The implementation of complex infinite dimensional models, specially the complicated mean restricted model for the scaled errors, will be computationally intensive in a multivariate setting like ours. The computational simplicity of the finite dimensional methods proposed in this article make them particularly suitable for multivariate problems.

In this paragraph, we discuss additional mixing issues that render infinite dimensional models, particularly the ones with non or semiconjugate priors on the component specific parameters (like our MLFA model), unsuitable for multivariate applications. There are two main types of MCMC algorithms for fitting infinite dimensional mixture models - conditional methods and marginal methods. In the conditional scheme, the mixture probabilities are sampled. The mixture labels are then updated independently, conditional on the mixture probabilities. The mixture probabilities in infinite dimensional mixture models can be stochastically ordered. For instance, mixture probabilities in a Dirichlet process mixture model satisfy $E(\widetilde{\pi}_k) > E(\widetilde{\pi}_{k+1})$ and $\Pr(\widetilde{\pi}_k > \widetilde{\pi}_{k+1}) > 0.5$ for all $k \in \mathbb{N}$. This imposes weak identifiability on the mixture labels resulting in a complicated model space comprising many local modes of varying importance. Different permutations of the mixture labels are not equivalent and exploration of the entire model space becomes important for valid inference. In high dimensional and large data settings it is difficult to achieve even by sophisticated MCMC algorithms with carefully designed label switching moves (Hastie, et al. 2013). The problem can be avoided with marginal methods (Neal, 2000) that integrate out the mixture probabilities and work with the resulting Polya urn scheme, rendering the mixture labels dependent but nonidentifiable. Unfortunately, such integration is possible only when conjugate priors are assigned to the component specific parameters. Typically for infinite dimensional models with non or semiconjugate priors on the component specific parameters, good mixing is thus difficult to achieve, particularly in complicated multivariate setup like ours.

Such issues also plague finite dimensional truncation based approximations to Dirichlet process mixture models where the mixture probabilities are constructed as $\widetilde{\pi}_k = V_k \prod_{\ell=1}^{k-1}(1 - V_\ell), V_\ell \sim \mathrm{Beta}(1, \alpha), k = 1, \ldots, (K-1)$, and $V_K = 1$ (Ishwaran and James, 2002) and the mixture components remain weakly identifiable.

On the contrary, the issues of mixing and convergence become much less important for finite mixture models with symmetric priors $(\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$ on the mixture probabilities. With $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ mixture components for the densities $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, respectively, the posterior is still multimodal but comprises $K_{\mathbf{X}}! \times K_{\boldsymbol{\epsilon}}!$ modal regions that are exact copies of each other. For inference on the overall density or any other functions of interest that are invariant to permutations of the mixture labels, it is only important that the MCMC sampler visits and explores at least one of the modal regions well and label switching (or the lack of it) does not present any problem (Geweke, 2007).

S.6.3 Model Selection and Model Averaging

As mentioned at the beginning of Section S.6, a major advantage of infinite dimensional mixture models is their ability to implicitly and semiautomatically perform model selection and model averaging. Properties of overfitted mixture models can be exploited to achieve the same in finite dimensional models with sufficiently large number of components. Recently Rousseau and Mengersen (2011) studied the asymptotic behavior of the posterior for overfitted mixture models with Dirichlet prior $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ on the mixture probabilities in a measurement error free set up and showed that the hyper parameter $(\alpha_1, \ldots, \alpha_k)$ strongly influences the way the posterior handles overfitting. In particular, when $\max_{k=1,\ldots,K} \alpha_k < L/2$, where L denotes the number of parameters specifying the component kernels, the posterior is asymptotically stable and concentrates in regions with empty redundant components. In this article, we chose symmetric Dirichlet priors $\mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$ on the mixture probabilities to model both the $f_{\mathbf{X}}$ and the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$. We set $\alpha_{\mathbf{X}} = \alpha_{\boldsymbol{\epsilon}} = 1$ so that the condition $\alpha/K < L/2$ is satisfied for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$. In simulation experiments reported in Section 6, the behavior of the posterior was similar to that observed by Rousseau and Mengersen (2011) in measurement error free set up. That is, when $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ were assigned sufficiently large values, the MCMC chain quickly reached a stable stage where the redundant components became empty. See Figure S.6 in the main article and Figure S.12 and S.13 in the Supplementary Material for illustrations, where, with some abuse of nomenclature, the k^{th} component is called empty if the associated mixture probability $\pi_k \leq 0.05$. Since such overfitted mixture models allow the number of nonempty mixture components to vary from one MCMC iteration to the next, model averaging is automatically achieved. MAP estimates of the numbers of mixture components required to approximate the target densities are given by the numbers of components which are visited the maximum number of times by the MCMC sampler, as in the case of infinite mixture models.

As discussed in the main paper, for the MIW method, when the measurement errors are conditionally heteroscedastic and the true covariance matrices are highly sparse, the strategy usually overestimates the number of non-empty mixture components required to approximate the target densities well. In these cases, the MIW method becomes highly numerically unstable and much larger sample sizes are required for the asymptotic results to hold. See Figure S.5 in the main article for an illustration. This may be regarded more as a limitation of the MIW method than a limitation of the adopted strategy to determine $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$. For the numerically more stable MLFA model, the asymptotic results are valid even for moderate sample sizes and such models are also more robust to overestimation of the number of nonempty clusters.

S.6.4 Model Flexibility

The proofs of the support results presented in Section 5 require that the number of mixture components of the corresponding mixture models be allowed to vary over the set of all positive integers. However, as the technical details of the proofs reveal, the use of mixture models with fixed finite number of components does not necessarily imply a compromise on the issue of flexibility. Indeed, a common recurring idea in the proofs of all these results, including those for the variance functions, is to show that any function coming from the target class can be approximated with any desired level of accuracy by the corresponding finite mixture models provided the models comprise sufficiently large number of mixture components and the function satisfies some fairly minimal regularly conditions. The requirement that the priors on the number of mixture components assign positive probability to all positive integers only helps us reach the final conclusions as immediate consequences. For any given data set of finite size, the number of mixture components required to approximate a target density will always be bounded above by the number of latent or observed variables generated by the target density. For most practical applications the required number would actually be much smaller than the number of variables generated by the target. Even if one applies mixture models that a-priori allow potentially infinitely many mixture components, the posterior will essentially concentrate on a finite set comprising moderately small positive integers. This means that for all practical purposes, solutions based on finite mixture models with fixed but sufficiently large number of mixture components will essentially be as robust as solutions based on their infinite or varying dimensional counterparts while at the same time being significantly less burdensome from a computational viewpoint. The requirement that the priors on the number of mixture components assign positive mass on *all* positive integers may thus be relegated to the requirement that the priors assign positive mass on sets of the form $\{1, \ldots, K\}$, where K is sufficiently large. Posterior computation for such models might be even much more intensive and complex requiring reversible jump moves. Since a mixture model with K components is at least as flexible as a model with $(K-1)$ components, properties of overfitted mixture models discussed in Section S.6.3 allow us to adopt a much simpler strategy. We can simply keep the number of mixture components fixed at sufficiently large values for all MCMC iterations. Carefully chosen priors for the mixture probabilities then result in a posterior that concentrates in regions favoring empty redundant components, essentially eliminating the need to assign any priors on the number of mixture components. We will still need some mechanism, preferably an automated and data adaptive one, to determine what values of K would be sufficiently large. This issue is discussed in the section on hyper-parameter choices in Section S.1.

The discussions of Section S.6 suggest that finite mixture models with sufficiently large number of mixture components and carefully chosen priors for the mixture probabilities can essentially retain the major advantages of infinite dimensional alternatives including flexibility, automated model averaging and model selection while at the same time being computationally much less burdensome, making them our preferred choice for complicated high dimensional problems.

S.7 Proofs of Theoretical Results of Section 5

S.7.1 Proof of Lemma 2

Proof of part 1 of Lemma 2 follows mostly by modifications of the results of Norets and Pelenis (2012). We present here only the proof of part 2 that requires additional modifications along the lines of Pelenis (2014) to accommodate the mean zero restriction on the density of the measurement errors. The first step is to construct finite mixture models of the form

$$f_m(\mathbf{z}|\boldsymbol{\theta}_m) = \sum_{k=1}^{m+2} \pi_{m,k} \ \mathrm{MVN}_p(\mathbf{z}|\boldsymbol{\mu}_{m,k}, \boldsymbol{\Sigma}_{m,k}) \quad \text{with} \quad \sum_{k=1}^{m+2} \pi_{m,k}\boldsymbol{\mu}_{m,k} = \mathbf{0}$$

that can approximate any given density f_0 that has mean zero and satisfies Conditions 1 with any desired level of accuracy. The continuity of $f_m(\cdot|\boldsymbol{\theta})$ implies that the KL distance between f_0 and f_m remains small on sufficiently small open neighborhoods around $\boldsymbol{\theta}_m$. Both the MIW and the MLFA priors assign positive probability to open neighborhoods around $\boldsymbol{\theta}_m$. The conclusion of part 2 of Lemma 2 follows since the prior probability of having $(m+2)$ mixture components is also positive for all $m \in \mathbb{N}$.

Lemma 5. *For any* $f_0 \in \widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}}$ *and* $\eta > 0$, *there exists* $\boldsymbol{\theta}_m$ *such that* $d_{KL}\{f_0(\cdot), f_m(\cdot|\boldsymbol{\theta}_m)\} < \eta$.

Proof. Let $\{A_{m,k}\}_{k=1}^m$ be adjacent cubes with side length h_m, and $A_{m,0} = \mathbb{R}^p - \cup_{k=1}^m A_{m,k}$ such that $h_m \downarrow 0$ but $\cup_{k=1}^m A_{m,k} \uparrow \mathbb{R}^p$ as $m \to \infty$. So $\{A_{m,k}\}_{k=1}^m$ becomes finer but $\cup_{k=1}^m A_{m,k}$ covers more of $\mathbb{R}^p$ as m increases. Additionally, let the partition be constructed in such a way that for all m sufficiently large, if $\boldsymbol{\epsilon} \in A_{m,0}$, then $C_r(\boldsymbol{\epsilon}) \cap A_{m,0}$ contains a hypercube $C_0(\boldsymbol{\epsilon})$ with side length $r/2$ and a vertex at $\boldsymbol{\epsilon}$; and if $\boldsymbol{\epsilon} \notin A_{m,0}$, then $C_r(\boldsymbol{\epsilon}) \cap (\mathbb{R}^p - A_{m,0})$ contains a hypercube $C_1(\boldsymbol{\epsilon})$ with side length $r/2$ and a vertex at $\boldsymbol{\epsilon}$. Consider the model

$$f_m(\mathbf{z}) \ = \ f_m(\mathbf{z}|\boldsymbol{\theta}_m) = \sum_{k=1}^{m+2} \pi_{m,k} \ \mathrm{MVN}_p(\mathbf{z}|\boldsymbol{\mu}_{m,k}, \boldsymbol{\Sigma}_{m,k}).$$

Set $\pi_{m,k} = \int_{A_{m,k}} f_0(\mathbf{z})d\mathbf{z}$ for $k = 1, 2, \ldots, m$ and $\pi_{m,k} = P_{f_0}(A_{m,0})/2 = \int_{A_{m,k}} f_0(\mathbf{z})d\mathbf{z}/2$ for $k = (m+1), (m+2)$. Then $\sum_{k=1}^{m+2} \pi_{m,k} = \int_{\mathbb{R}^p} f_0(\mathbf{z})d\mathbf{z} = 1$. Define $g(\mathbf{d}) = \sum_{k=1}^m \pi_{m,k}(\mathbf{c}_{m,k} + \mathbf{d}) + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z}$, where $\mathbf{c}_{m,k}$ is the center of $A_{m,k}$ for $k = 1, 2, \ldots, m$.

$$\begin{aligned} g(h_m\mathbf{1}_p/2) &= \sum_{k=1}^m \pi_{m,k}(\mathbf{c}_{m,k} + h_m\mathbf{1}_p/2) + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} \\ &= \sum_{k=1}^m \int_{A_{m,k}} (\mathbf{c}_{m,k} + h_m\mathbf{1}_p/2) f_0(\mathbf{z})d\mathbf{z} + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} \\ &\geq \sum_{k=1}^m \int_{A_{m,k}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} = \int_{\mathbb{R}^p} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} = \mathbf{0}. \end{aligned}$$

Similarly $g(-h_m\mathbf{1}_p/2) \leq 0$. Since $g(\cdot)$ is continuous, there exists $\mathbf{d}_m \in [-h_m/2, h_{m/2}]^p$ such that $g(\mathbf{d}_m) = \mathbf{0}$. Set $\boldsymbol{\mu}_{m,k} = (\mathbf{c}_{m,k} + \mathbf{d}_m)$ for $k = 1, 2, \ldots, m$. Also set $\boldsymbol{\mu}_{m,m+1} =$

$2\int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z}/\int_{A_{m,0}} f_0(\mathbf{z})d\mathbf{z}$ and $\boldsymbol{\mu}_{m,m+2} = \mathbf{0}$ when $\int_{A_{m,0}} f_0(\mathbf{z})d\mathbf{z} > 0$, and $\boldsymbol{\mu}_{m,0} = \mathbf{0}$ otherwise. Then $\sum_{k=1}^{m+2} \pi_{m,k}\boldsymbol{\mu}_{m,k} = g(\mathbf{d}_m) = \mathbf{0}$. Also set $\boldsymbol{\Sigma}_{m,k} = \sigma_m^2 \mathrm{I}_p$ for $k = 1, 2, \ldots, m$ with $\sigma_m \to 0$, and $\Sigma_{m,m+1} = \Sigma_{m,m+2} = \sigma_0^2 \mathrm{I}_p$.

Consider a sequence $\{\delta_m\}_{m=1}^{\infty}$ satisfying $\delta_m > 6p^{1/2}h_m$ and $\delta_m \to 0$. Fix $\boldsymbol{\epsilon} \in \mathbb{R}^p$. Define $C_{\delta_m}(\boldsymbol{\epsilon}) = [\boldsymbol{\epsilon} - \delta_m \mathbf{1}_p/2, \boldsymbol{\epsilon} + \delta_m \mathbf{1}_p/2]$. For m sufficiently large $C_{\delta_m}(\boldsymbol{\epsilon}) \subseteq \cup_{k=1}^{m} A_{m,k}$, $C_{\delta_m}(\boldsymbol{\epsilon}) \cap A_{m,0} = \phi$ and the set $\{k : 1 \leq k \leq m, A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})\}$ is non-empty. For $k = 1, \ldots, m$, when $A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})$, $\pi_{m,k} \geq \inf_{\mathbf{z} \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) h_m^p$. Therefore,

$$\begin{aligned} f_m(\boldsymbol{\epsilon}) &\geq \sum_{\{k:1\leq k\leq m, A_{m,k}\subset C_{\delta_m}(\boldsymbol{\epsilon})\}} \pi_{m,k} \; \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\ &\geq \inf_{z\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \sum_{\{k:A_{m,k}\subset C_{\delta_m}(\boldsymbol{\epsilon})\}} h_m^p \; \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{c}_{m,k} + \mathbf{d}_m, \sigma_m^2 \mathrm{I}_p) \\ &\geq \inf_{z\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \; \left\{1 - \frac{6p^{3/2}h_m\delta_m^{p-1}}{(2\pi)^{p/2}\sigma_m^p} - \frac{8p\sigma_m}{(2\pi)^{1/2}\delta_m}\right\}, \end{aligned}$$

where the last step follows from Lemma 1 and Lemma 2 of Norets and Pelenis (2012). Let h_m, δ_m, σ_m further satisfy $h_m/\sigma_m^p \to 0, \sigma_m/\delta_m \to 0$. Then for any $\eta > 0$ there exists an M_1 large enough such that for all $m > M_1$

$$f_m(\boldsymbol{\epsilon}) \geq \inf_{\mathbf{z}\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \cdot (1-\eta).$$

Without loss of generality, we may assume $f_0(\boldsymbol{\epsilon}) > 0$. Since $f_0(\cdot)$ is continuous and $\delta_m \to 0$, there also exists an M_2 such that for all $m > M_2$ we have $\inf_{\mathbf{z}\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) > 0$ and

$$\frac{f_0(\boldsymbol{\epsilon})}{\inf_{\mathbf{z}\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z})} \leq (1+\eta).$$

Therefore, for all $m > \max\{M_1, M_2\}$, we have

$$1 \leq \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon})}\right\} \leq \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{\inf_{z\in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(z)\cdot(1-\eta)}\right\} \leq \frac{(1+\eta)}{(1-\eta)}.$$

Thus, $\log\max\{1, f_0(\boldsymbol{\epsilon})/f_m(\boldsymbol{\epsilon})\} \to 0$ as $m \to \infty$. Pointwise convergence is thus established. Next, we will find an integrable upper bound for $\log\max\{1, f_0(\boldsymbol{\epsilon})/f_m(\boldsymbol{\epsilon})\}$.

For point wise convergence we can assume $\boldsymbol{\epsilon} \notin A_{m,0}$ for sufficiently large m. But to find integrable upper bound, we have to consider both the cases $\boldsymbol{\epsilon} \in A_{m,0}$ and $\boldsymbol{\epsilon} \notin A_{m,0}$. When $\boldsymbol{\epsilon} \in A_{m,0}$, we have $\mathrm{P}_{f_0}(A_{m,0}) = \int_{A_{m,0}} f_0(\mathbf{z})d\mathbf{z} \geq \int_{A_{m,0}\cap C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})d\mathbf{z} \geq \lambda\{A_{m,0} \cap C_r(\boldsymbol{\epsilon})\} \inf_{\mathbf{z}\in A_{m,0}\cap C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \geq (r/2)^p \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})$, since $\lambda\{A_{m,0} \cap C_r(\boldsymbol{\epsilon})\} \geq \lambda\{C_0(\boldsymbol{\epsilon})\} \geq (r/2)^p$. Using part 4 of Conditions 1 and Lemma 1 and Lemma 2 of Norets and Pelenis (2012) again, if $\boldsymbol{\epsilon} \notin A_{m,0}$, for m sufficiently large

$$\sum_{\{k:A_{m,k}\subset C_r(\boldsymbol{\epsilon})\}} h_m^p \; \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \geq \sum_{\{k:A_{m,k}\subset C_1(\boldsymbol{\epsilon})\}} h_m^p \; \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p)$$

$$
\begin{aligned}
&\geq \int_{C_1(\boldsymbol{\epsilon})} \mathrm{MVN}_p(\mathbf{z}|\boldsymbol{\epsilon}, \sigma_m^2 \mathrm{I}_p) d\mathbf{z} - \frac{3p^{3/2}(r/2)^{p-1}h_m}{(2\pi)^{p/2}\sigma_m^p} \\
&\geq \left\{ \frac{1}{2^p} - \frac{8p\sigma_m}{2^p(2\pi)^{1/2}r} - \frac{3p^{3/2}h_m r^{p-1}}{2^{p-1}(2\pi)^{p/2}\sigma_m^p} \right\} \geq \frac{1}{2^{p+1}},
\end{aligned}
$$

This implies

$$
\begin{aligned}
f_m(\boldsymbol{\epsilon}) \;&=\; \sum_{k=1}^{m} P_{f_0}(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) + \sum_{k=m+1}^{m+2} (1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_0^2 \mathrm{I}_p) \\
&\geq\; \sum_{k=1}^{m} P_{f_0}(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) + (1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) \\
&\geq\; \{1 - 1(\boldsymbol{\epsilon} \in A_{m,0})\} \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \sum_{\{k: A_{m,k} \subset C_r(\boldsymbol{\epsilon})\}} \lambda(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\
&\quad + 1(\boldsymbol{\epsilon} \in A_{m,0})(1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) \\
&\geq\; (1/2)\{1 - 1(\boldsymbol{\epsilon} \in A_{m,0})\} \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \\
&\quad + 1(\boldsymbol{\epsilon} \in A_{m,0}) \ (1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \\
&\geq\; (1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}).
\end{aligned}
$$

The last step followed by choosing σ_0^2 large enough so that $(r/2)^p \sup_{\boldsymbol{\epsilon} \in \mathbb{R}^p} \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) < (r/2)^p \ \sigma_0^{-p} < 2^{-(p+1)} < 1$. Therefore,

$$
\begin{aligned}
&\log\max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon})}\right\} \leq \log\max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p) \ \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})}\right\} \\
&\quad \leq \log\left[\frac{1}{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p)} \max\left\{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p), \frac{f_0(\boldsymbol{\epsilon})}{\inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})}\right\}\right] \\
&\quad \leq -\log\left\{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_0^2 \mathrm{I}_p)\right\} + \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{\inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})}\right\}.
\end{aligned}
$$

The first and the second terms are integrable by part 2 and part 3 of Conditions 1, respectively. Since $\int f_0(\boldsymbol{\epsilon}) \log\{f_0 \boldsymbol{\epsilon}/f_m(\boldsymbol{\epsilon})\} d\boldsymbol{\epsilon} \leq \int f_0(\boldsymbol{\epsilon}) \log\max\{1, f_0 \boldsymbol{\epsilon}/f_m(\boldsymbol{\epsilon})\} d\boldsymbol{\epsilon}$, the proof of Lemma 5 is completed applying dominated convergence theorem (DCT). ☐

Let $\eta > 0$ be given. According to Lemma 5, there exists $\boldsymbol{\theta}_m^\star = (\boldsymbol{\pi}_{1:(m+2)}^\star, \boldsymbol{\mu}_{1:(m+2)}^\star, \boldsymbol{\Sigma}_{1:(m+2)}^\star)$ with $\boldsymbol{\Sigma}_k^\star = \sigma_m^{2\star} \mathrm{I}_p$ for $k = 1, \dots, m$ and $\boldsymbol{\Sigma}_k^\star = \sigma_0^{2\star} \mathrm{I}_p$ for $k = (m+1), (m+2)$ such that $d_{KL}\{f_0(\cdot), f_m(\cdot|\boldsymbol{\theta}_m^\star)\} < \eta/2$. We have, for any $\boldsymbol{\theta}_m$,

$$
\int f_0(\boldsymbol{\epsilon}) \ \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)}\right\} d\boldsymbol{\epsilon} = \int f_0(\boldsymbol{\epsilon}) \ \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m^\star)}\right\} d\boldsymbol{\epsilon} + \int f_0(\boldsymbol{\epsilon}) \ \log\left\{\frac{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)}\right\} d\boldsymbol{\epsilon}.
$$

Let the second term in the above expression be denoted by $g(\boldsymbol{\theta}_m)$. The priors puts positive mass on arbitrarily small open neighborhoods around $\boldsymbol{\theta}_m^\star$. The result will follow if there exists an open neighborhood $\mathcal{N}(\boldsymbol{\theta}_m^\star)$ around $\boldsymbol{\theta}_m^\star$ such that $\sup_{\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)} g(\boldsymbol{\theta}_m) < \eta/2$.

Since $g(\boldsymbol{\theta}_m^\star)=0$, it suffices to show that the function $g(\boldsymbol{\theta}_m)$ is continuous at $\boldsymbol{\theta}_m^\star$. Now $g(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_m^\star$ if for every sequence $\{\boldsymbol{\theta}_{m,n}\}_{n=1}^\infty$ with $\boldsymbol{\theta}_{m,n}\to\boldsymbol{\theta}_m^\star$, we have $g(\boldsymbol{\theta}_{m,n})\to g(\boldsymbol{\theta}_m^\star)$. For all $\boldsymbol{\epsilon}\in\mathbb{R}^p$, we have $\log\{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_{m,n}^\star)/f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)\}\to 0$ as $\boldsymbol{\theta}_{m,n}\to\boldsymbol{\theta}_m^\star$. Continuity of $g(\boldsymbol{\theta}_m)$ at $\boldsymbol{\theta}_m^\star$ will follow from DCT if we can show that $|f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m^\star)/f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_{m,n})|$ has an integrable with respect to f_0 upper bound.

Since $\boldsymbol{\theta}_{m,n}\to\boldsymbol{\theta}_m^\star$, for any arbitrarily small open neighborhood $\mathcal{N}(\boldsymbol{\theta}_m^\star)$ around $\boldsymbol{\theta}_m^\star$, we must have $\boldsymbol{\theta}_{m,n}\in\mathcal{N}(\boldsymbol{\theta}_m^\star)$ for all n sufficiently large. Let $\boldsymbol{\theta}_m=(\boldsymbol{\pi}_{1:(m+2)},\boldsymbol{\mu}_{1:(m+2)},\boldsymbol{\Sigma}_{1:(m+2)})\in\mathcal{N}(\boldsymbol{\theta}_m^\star)$. Since the eigenvalues of a real symmetric matrix depend continuously on the matrix, we must have $(\lambda_1(\boldsymbol{\Sigma}_k),\lambda_p(\boldsymbol{\Sigma}_k))\subset(\underline{\sigma}_m^{2\star},\overline{\sigma}_m^{2\star})$ for $k=1,\dots,m$ and $(\lambda_1(\boldsymbol{\Sigma}_k),\lambda_p(\boldsymbol{\Sigma}_k))\subset(\underline{\sigma}_0^{2\star},\overline{\sigma}_0^{2\star})$ for $k=(m+1),(m+2)$, where $\underline{\sigma}_m^{2\star}<\sigma_m^{2\star}<\overline{\sigma}_m^{2\star}$ and $\underline{\sigma}_0^{2\star}<\sigma_0^{2\star}<\overline{\sigma}_0^{2\star}$. Let $\underline{\sigma}^{2\star}=\min\{\underline{\sigma}_m^{2\star},\underline{\sigma}_0^{2\star}\}$ and $\overline{\sigma}^{2\star}=\max\{\overline{\sigma}_m^{2\star},\overline{\sigma}_0^{2\star}\}$. Then $(\lambda_1(\boldsymbol{\Sigma}_k),\lambda_p(\boldsymbol{\Sigma}_k))\subset(\underline{\sigma}^{2\star},\overline{\sigma}^{2\star})$ for $k=1,\dots,(m+2)$. Similarly, for some finite $\mu^\star$, we must have $\boldsymbol{\mu}_{m,k}\in(-\mu^\star\mathbf{1}_p,\mu^\star\mathbf{1}_p)=\mathcal{N}_{\mu^\star}$ for $k=1,\dots,(m+2)$. For any real positive definite matrix $\boldsymbol{\Sigma}$, we have $\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\leq\lambda_1^{-1}(\boldsymbol{\Sigma})\left\|\mathbf{z}\right\|^2$. Therefore, for any $\boldsymbol{\epsilon}\in\mathbb{R}^p$ and for all $k=1,\dots,(m+2)$, we must have $(\boldsymbol{\epsilon}-\boldsymbol{\mu}_{m,k})^{\mathrm{T}}\boldsymbol{\Sigma}_{m,k}^{-1}(\boldsymbol{\epsilon}-\boldsymbol{\mu}_{m,k})\leq\underline{\sigma}^{-2\star}\{1(\boldsymbol{\epsilon}\in\mathcal{N}_{\mu^\star})2^p\mu^{\star p}+1(\boldsymbol{\epsilon}\notin\mathcal{N}_{\mu^\star})\left\|\boldsymbol{\epsilon}+\mathrm{sign}(\boldsymbol{\epsilon})\mu^\star\right\|^2\}$, where $\mathrm{sign}(\boldsymbol{\epsilon})=\{\mathrm{sign}(\epsilon_1),\dots,\mathrm{sign}(\epsilon_p)\}^{\mathrm{T}}$. Therefore, for any $\boldsymbol{\theta}_m\in\mathcal{N}(\boldsymbol{\theta}_m^\star)$, we have

$$\begin{aligned}[1(\boldsymbol{\epsilon}\in\mathcal{N}_{\mu^\star})\mathrm{MVN}_p(2\mu^\star\mathbf{1}_p|\mathbf{0},\underline{\sigma}^{2\star}\mathrm{I}_p)+1(\boldsymbol{\epsilon}\notin\mathcal{N}_{\mu^\star})\mathrm{MVN}_p\{\boldsymbol{\epsilon}+\mathrm{sign}(\boldsymbol{\epsilon})\mu^\star|\mathbf{0},\underline{\sigma}^{2\star}\mathrm{I}_p\}]/\overline{\sigma}^\star&\\ \leq f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)\leq 1/\underline{\sigma}^\star.&\end{aligned}$$

The upper bound is a constant and the logarithm of the lower bound is integrable since, by part 2 of Conditions 1, the second order moments of $\boldsymbol{\epsilon}$ exist. An f_0 integrable upper bound for the function $\sup_{\boldsymbol{\theta}_m\in\mathcal{N}(\boldsymbol{\theta}_m^\star)}|f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)|$ thus exists. Finally, DCT applies because

$$\begin{aligned}\int f_0(\boldsymbol{\epsilon})\ \left|\log\left\{\frac{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_{m,n})}\right\}\right|d\boldsymbol{\epsilon} &\leq \sup_{\boldsymbol{\theta}_m\in\mathcal{N}(\boldsymbol{\theta}_m^\star)}\int f_0(\boldsymbol{\epsilon})\ \left|\log\left\{\frac{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)}\right\}\right|d\boldsymbol{\epsilon}\\ &\leq 2\sup_{\boldsymbol{\theta}_m\in\mathcal{N}(\boldsymbol{\theta}_m^\star)}\int f_0(\boldsymbol{\epsilon})\ \left|f_m(\boldsymbol{\epsilon}|\boldsymbol{\theta}_m)\right|d\boldsymbol{\epsilon}.\end{aligned}$$

The conclusion of part 2 of Lemma 2 follows since the prior probability of having $(m+2)$ mixture components is positive for all $m\in\mathbb{N}$.

S.7.2 Proof of Lemma 3

Given q, let Π_q denote a prior on $\mathbb{N}_q=\{q+1,q+2,\dots\}$ such that $\Pi_q(J)>0\ \forall J\in\mathbb{N}_q$. Let $||\cdot||_2$ denote the Euclidean norm. Let $\mathbb{R}^+=(0,\infty)$. Given $J\sim\Pi_q$, also let $\Pi_{\beta|J}$ be a prior on $\mathbb{R}^{+J}$ such that $\Pi_{\beta|J}\{N_\delta(\boldsymbol{\beta}_0)\}>0$ for any $\delta>0$ and any $\boldsymbol{\beta}_0\in\mathbb{R}^J$, where $N_\delta(\boldsymbol{\beta}_0)=\{\boldsymbol{\beta}:\boldsymbol{\beta}\in\mathbb{R}^{+J},||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2<\delta\}$. Define $\mathcal{S}_{q,J}=\{v_s:v_s=\mathbf{B}_{q,J}\boldsymbol{\beta}=\sum_{j=1}^J b_{q,j}\beta_j$ for some $\boldsymbol{\beta}\in\mathbb{R}^{+J}\}$. Then $\Pi_{\mathbf{V}}=\Pi_q\times\Pi_{\beta|J}$ is the induced prior on $\mathcal{S}_q=\cup_{J=q+1}^\infty\mathcal{S}_{q,J}$.

Define $\psi(v_0,h)=\sup_{X,X'\in[A,B],|X-X'|\leq h}|v_0(X)-v_0(X')|$. Let $\lfloor\alpha\rfloor=\min\{n:n\in\mathbb{N},n\geq\alpha\}$. For any X, (i) $b_{q,j}(X)\geq 0\ \forall j$, (ii) $\sum_{j=1}^J b_{q,j}(X)=1$, (iii) $b_{q,j}$ is positive only inside the interval $[t_j, t_{j+q+1}]$, and (iv) for $j \in \{(q+1), (q+2), \ldots, (q+K)\}$, for any $X \in (t_j, t_{j+1})$, only $(q+1)$ B-splines $b_{q,j-q}(X), b_{q,j-q+1}(X), \ldots, b_{q,j}(X)$ are positive. Using these local support properties of B-splines, the results on page 147 of de Boor (2000) can be modified to show that, for any $v_0 \in \mathcal{C}_+[A,B]$,

$$\inf_{v_s \in \mathcal{S}_{q,J}} ||v_0 - v_s||_\infty \le \lfloor (q+1)/2 \rfloor \ \psi(v_0, \Delta_{\max}) \to 0 \ \text{ as } \Delta_{\max} \to 0.$$

Also, if $q \ge (\alpha - 1)$, we can modify the results on page 149 of de Boor (2000) to show that, for any $v \in \mathcal{C}_+^\alpha[A,B]$,

$$\inf_{v_s \in \mathcal{S}_{q,J}} ||v_0 - v_s||_\infty \ \ \le \ \ c(q)c(q-1)\ldots c(q-\alpha_0+1) \ ||v_0^{(\alpha_0)}||_\infty \ \Delta_{\max}^{\alpha_0},$$

where $c(q) = \lfloor (q+1)/2 \rfloor$. For any two functions g_1 and g_2, $\sup|g_1 g_2| \le \sup|g_1| \sup|g_2|$. Taking $g_1(X,X') = \{v_0^{(\alpha_0)}(X) - v_0^{(\alpha_0)}(X')\}/(X-X')^{(\alpha-\alpha_0)}$ and $g_2(X,X') = (X-X')^{(\alpha-\alpha_0)}$, we have $||v_0^{(\alpha_0)}||_\infty \le ||v_0||_\alpha (B-A)^{(\alpha-\alpha_0)}$. Therefore,

$$\inf_{v_s \in \mathcal{S}_{q,J}} ||v_0 - v_s||_\infty \ \ \le \ \ c(q,\alpha_0) \ (B-A)^{(\alpha-\alpha_0)} \ ||v_0||_\alpha \ \Delta_{\max}^{\alpha_0}.$$

Furthermore, when the knot points $\{t_{q+1+j}\}_{j=0}^K$ are equidistant

$$\inf_{v_s \in \mathcal{S}_{q,J}} ||v_0 - v_s||_\infty \le c(q,\alpha_0) ||v_0^{(\alpha)}||_\infty \frac{(B-A)^\alpha}{K^{\alpha_0}} \le c(q,\alpha) ||v_0||_\alpha K^{-\alpha}.$$

Given any $v_0 \in C_+[A,B]$(or $C_+^\alpha[A,B]$) and $\delta > 0$, find $J \in \mathbb{N}_q$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^{+J}$ such that $||v_0 - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty = \inf_{v_s \in \mathcal{S}_{q,J}} ||v_0 - v_s||_\infty < \delta/2$. Next consider a neighborhood $N_\eta(\boldsymbol{\beta}_0)$ such that for any $\boldsymbol{\beta} \in N_\eta(\boldsymbol{\beta}_0)$, we have $||\mathbf{B}_{q,J}\boldsymbol{\beta} - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty < \delta/2$. Then for any $\boldsymbol{\beta} \in N_\eta(\boldsymbol{\beta}_0)$, we have $||\mathbf{B}_{q,J}\boldsymbol{\beta} - v_0||_\infty \le ||\mathbf{B}_{q,J}\boldsymbol{\beta} - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty + ||\mathbf{B}_{q,J}\boldsymbol{\beta}_0 - v_0||_\infty < \delta$. Also $\Pi_{\mathbf{V}}(||v - v_0||_\infty < \delta) \ge \Pi_q(J) \ \Pi_{\beta|J}\{N_\eta(\boldsymbol{\beta}_0)\} > 0$. Proof of Lemma 3 then follows as a special case taking $\boldsymbol{\beta} = \exp(\boldsymbol{\xi})$ and taking Π_q and $\Pi_{\beta|J}$ to be the priors on J and $\boldsymbol{\beta}$ induced by $P_0(K)$ and $P_0(\boldsymbol{\xi}|K, \sigma_\xi^2)$, respectively.

S.7.3 Proof of Lemma 4

We first prove some additional lemmas to used in the proof of Lemma 4.

Lemma 6. $\Pi_{\mathbf{V}}(||v - v_0||_\infty < \delta) > 0 \ \forall \delta > 0$ *implies that* $\Pi_{\mathbf{V}}(||g \circ v - g \circ v_0||_\infty < \delta) > 0 \ \forall \delta > 0$ *for every continuous function* $g : \mathbb{R} \to \mathbb{R}$.

Proof. Let $v : [A,B] \to [C_1, D_1]$ and $v_0 : [A,B] \to [C_2, D_2]$. Then $(v - v_0) : [A,B] \to [C_1 - D_2, D_1 - C_2] = [C,D]$, say. Then $g : [C,D] \to \mathbb{R}$ is a uniformly continuous function. Therefore, given any $\delta > 0$, there exists a $\eta > 0$ such that $|g(Z_1) - g(Z_2)| < \delta$ whenever $|Z_1 - Z_2| < \eta$. Now let $||v - v_0||_\infty = \sup_{X \in [A,B]} |v(X) - v_0(X)| < \eta$. This implies, for all $X \in [A,B]$, $|v(X) - v_0(X)| < \eta$. Therefore, for all $X \in [A,B]$, $|g\{v(X)\} - g\{v_0(X)\}| < \delta$, and hence $||g \circ v - g \circ v_0||_\infty \le \delta$. Hence the proof. $\square$

Corollary 1. *In particular, taking* $g(Z) = Z^{1/2}\ \forall Z > 0$ *and* $g(Z) = 0$ *otherwise, we have* $\Pi_{\mathbf{V}}(||v^{1/2} - v_0^{1/2}||_\infty < \delta) = \Pi_{\mathbf{V}}(||s - s_0||_\infty < \delta) > 0\ \forall \delta > 0$ *for all* $v_0 \in \mathcal{C}_+[A, B]$*(or* $\mathcal{C}_+^\alpha[A, B]$*).*

Let $P_{\boldsymbol{\epsilon},K}\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) | \boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}\} = \sum_{k=1}^K \pi_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\delta_{\boldsymbol{\theta}}$ denotes a point mass at $\boldsymbol{\theta}$. We have, with the the hyper-parameters implicit, $P_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) = P_{0\pi}(\boldsymbol{\pi}_{1:K}) P_{0\mu}(\boldsymbol{\mu}_{1:K} | \boldsymbol{\pi}_{1:K}) P_{0\Sigma}(\boldsymbol{\Sigma}_{1:K})$. Denoting $P_{\boldsymbol{\epsilon},K}\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) | \boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}\}$ simply by $P_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let c be a generic for constants that are not of direct interest. For any square matrix $\mathbf{A}$ of order p, let $\lambda_1(\mathbf{A}) \leq \cdots \leq \lambda_p(\mathbf{A})$ denote the ordered eigenvalues of $\mathbf{A}$. The following lemma proves some properties of $P_{\boldsymbol{\epsilon},K}$ and $f_{\boldsymbol{\epsilon}}$.

Lemma 7. *1.* $\int \|\boldsymbol{\mu}\|_2^2 \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.* *2.* $\int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.*
3. $\int |\boldsymbol{\Sigma}|^{-1/2} \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.*

Proof. 1. The prior $P_{0\mu}(\boldsymbol{\mu}_{1:K} | \boldsymbol{\pi}_{1:K})$ is of the form (15), that is, $P_{0\mu}(\boldsymbol{\mu}_{1:K} | \boldsymbol{\pi}_{1:K}) = \mathrm{MVN}_{Kp}(\mathbf{0}, \boldsymbol{\Sigma}^0 - \boldsymbol{\Sigma}_{1,R}^0 \boldsymbol{\Sigma}_{R,R}^{-1} \boldsymbol{\Sigma}_{R,1}^0)$, where $\boldsymbol{\Sigma}^0$ is a $Kp \times Kp$ block-diagonal matrix independent of $\boldsymbol{\pi}_{1:K}$, all k principal blocks of order $p \times p$ being $\boldsymbol{\Sigma}_0$. The matrix $\boldsymbol{\Sigma}_{1,R}^0 \boldsymbol{\Sigma}_{R,R}^{-1} \boldsymbol{\Sigma}_{R,1}^0$ depends on $\boldsymbol{\pi}_{1:K}$ and is nonnegative definite so that its diagonal elements are all nonnegative. Let $\boldsymbol{\Sigma}_0 = ((\sigma_{0,ij}))$ and $\boldsymbol{\Sigma}_{1,R}^0 \boldsymbol{\Sigma}_{R,R}^{-1} \boldsymbol{\Sigma}_{R,1}^0 = ((\sigma_{R,ij}))$. Then, $\int \|\boldsymbol{\mu}_k\|_2^2 \, dP_{0\mu}(\boldsymbol{\mu}_{1:K} | \boldsymbol{\pi}_{1:K}) = \left\{\sum_{j=1}^p \sigma_{0,jj} - \sum_{j=(k-1)p+1}^{kp} \sigma_{R,jj}\right\} \leq \sum_{j=1}^p \sigma_{0,jj} = \mathrm{trace}(\boldsymbol{\Sigma}_0)$. Therefore,

$$\int \int \|\boldsymbol{\mu}\|_2^2 \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) = \sum_{k=1}^K \int \pi_k \|\boldsymbol{\mu}_k\|_2^2 \, dP_{0\mu}(\boldsymbol{\mu}_{1:K} | \boldsymbol{\pi}_{1:K}) dP_{0\pi}(\boldsymbol{\pi}_{1:K})$$
$$\leq \mathrm{trace}(\boldsymbol{\Sigma}_0) < \infty.$$

2. We have $\int \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) = \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{0\Sigma}(\boldsymbol{\Sigma})$.

When $\boldsymbol{\Sigma} \sim \mathrm{IW}_p(\nu_0, \boldsymbol{\Psi}_0)$, we have $\boldsymbol{\Psi}_0^{-1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Psi}_0^{-1/2} \sim \mathrm{W}_p(\nu_0, \mathrm{I})$ and $\mathrm{trace}(\boldsymbol{\Psi}_0^{-1} \boldsymbol{\Sigma}^{-1}) = \mathrm{trace}(\boldsymbol{\Psi}_0^{-1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Psi}_0^{-1/2}) \sim \chi^2_{p\nu_0}$. Here $\mathrm{W}_p(\nu, \boldsymbol{\Psi})$ denotes a Wishart distribution with degrees of freedom ν and mean $\nu\boldsymbol{\Psi}$. For any two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\lambda_1(\mathbf{A}) \mathrm{trace}(\mathbf{B}) \leq \mathrm{trace}(\mathbf{AB}) \leq \lambda_p(\mathbf{A}) \mathrm{trace}(\mathbf{B})$. Therefore, $\lambda_1(\boldsymbol{\Psi}_0^{-1}) E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} \leq E\{\mathrm{trace}(\boldsymbol{\Psi}_0^{-1} \boldsymbol{\Sigma}^{-1})\} = p\nu_0$. Hence, $\int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{0\Sigma}(\boldsymbol{\Sigma}) = E\lambda_p(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} < \infty$.

When $\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ with $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, we have $\mathrm{trace}(\boldsymbol{\Sigma}^{-1}) = \mathrm{trace}\{\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma} (\mathrm{I}_p + \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{\Omega}^{-1}\} \leq \mathrm{trace}(\boldsymbol{\Omega}^{-1}) = \sum_{j=1}^p \sigma_j^{-2}$, where $\boldsymbol{\Gamma}$ is a $p \times p$ matrix satisfying $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$. Thus, $\int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = E\lambda_p(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} \leq \sum_{j=1}^p E\sigma_j^{-2} < \infty$ whenever $\sigma_j^2 \sim \text{Inv-Ga}(a, b)$ with $a > 1$.

3. When $\boldsymbol{\Sigma} \sim \mathrm{IW}_p(\nu_0, \boldsymbol{\Psi}_0)$, we have $\lambda_1^{p/2}(\boldsymbol{\Psi}_0^{-1}) E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\}^{p/2} \leq E\{\mathrm{trace}(\boldsymbol{\Psi}_0^{-1} \boldsymbol{\Sigma}^{-1})\}^{p/2} < \infty$. Hence, $\int |\boldsymbol{\Sigma}|^{-1/2} \, dP_{0\Sigma}(\boldsymbol{\Sigma}) = \int \prod_{j=1}^p \lambda_j^{1/2}(\boldsymbol{\Sigma}^{-1}) dP_{0\Sigma}(\boldsymbol{\Sigma}) \leq \int \lambda_p^{p/2}(\boldsymbol{\Sigma}^{-1}) dP_{0\Sigma}(\boldsymbol{\Sigma}) = E\lambda_p^{p/2}(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\}^{p/2} < \infty$.

For any two positive semidefinite matrix $\mathbf{A}$ and $\mathbf{B}$, we have $|\mathbf{A} + \mathbf{B}| \geq |\mathbf{A}|$. Therefore, when $\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$, we have $\int |\boldsymbol{\Sigma}|^{-1/2} \, dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) \leq \int |\boldsymbol{\Omega}|^{-1/2} \, dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = \int \prod_{j=1}^p \sigma_j^{-1} dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = \prod_{j=1}^p E\sigma_j^{-1} < \infty$, whenever $\sigma_j^2 \sim \text{Inv-Ga}(a, b)$ independently. □

The following lemma proves a property of $f_{\boldsymbol{\epsilon}} = \int\int f_{c\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_0(K)$. Here $P_0(K)$ denotes the prior on K, the number of mixture components.

Lemma 8. *Let* $f_{0\boldsymbol{\epsilon}} \in \widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}}$ *and* $f_{\boldsymbol{\epsilon}} \sim \Pi_{\boldsymbol{\epsilon}}$ *and* $\mathbf{D}(\boldsymbol{\tau}) = diag(\tau_1, \tau_2, \ldots, \tau_p)$. *Then*

$$\lim_{\boldsymbol{\tau}\to\mathbf{1}} \int f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \; log \left[\frac{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}}\right] \; d\boldsymbol{\epsilon} = 0.$$

Proof. We have $|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{c\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\} \to f_{c\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ as $\boldsymbol{\tau} \to \mathbf{1}$. Since $\boldsymbol{\tau} \to \mathbf{1}$, without loss of generality, we may assume $|\mathbf{D}(\boldsymbol{\tau})| > 1/2$. Define $c = \int |\boldsymbol{\Sigma}|^{-1/2} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $c < \infty$. Also $\int |\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{c\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}|\boldsymbol{\theta}\} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le \int 2(2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < 2c < \infty$. Applying DCT, $|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\} \to f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ as $\boldsymbol{\tau} \to \mathbf{1}$. Therefore, for any $\boldsymbol{\epsilon} \in \mathbb{R}$,

$$\log\left[\frac{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}}\right] \to 0 \quad \text{as } \boldsymbol{\tau} \to \mathbf{1}.$$

To find an integrable with respect to $f_{0\boldsymbol{\epsilon}}$ upper bound for $\log\left[|\mathbf{D}(\boldsymbol{\tau})| f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})/f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}\right]$, we use Lemma 7. To do so, we can ignore the prior $P_0(K)$ since the upper bounds obtained in Lemma 7 do not depend on the specific choice of K. We have, using part 3 of Lemma 7,

$$\begin{aligned}\int |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \le \int |\boldsymbol{\Sigma}|^{-1/2} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le c.\end{aligned}$$

Since $\boldsymbol{\tau} \to \mathbf{1}$, without loss of generality we may also assume $\tau_k < 2$ for all k. Therefore,

$$\begin{aligned}
&|\log\, f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}| \\
&\quad \le \log(2\pi)^{p/2} + \left|\log \int |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right| \\
&\quad \le \log(2\pi)^{p/2} + |\log\, c| \\
&\qquad\qquad - \log \int c^{-1} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\quad \le \log\{c(2\pi)^{p/2}\} + |\log\, c| \\
&\qquad\qquad + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\int \{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\quad \le \log\{c(2\pi)^{p/2}\} + |\log\, c| \\
&\qquad\qquad + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\int \|\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} - \boldsymbol{\mu}\|_2^2\, \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\quad \le \log\{c(2\pi)^{p/2}\} + |\log\, c| \\
&\qquad\qquad + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \int \{\|\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\|_2^2 + \|\boldsymbol{\mu}\|_2^2\}\lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\quad \le \log\{c(2\pi)^{p/2}\} + |\log\, c| + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}$$

$$+ \|2\boldsymbol{\epsilon}\|_2^2 \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) + \int \|\boldsymbol{\mu}\|_2^2 \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

where the third step followed from application of Jensen's inequality on $g(Z) = -\log Z$. The regularity assumptions on $f_{0\boldsymbol{\epsilon}}$ and Lemma 7 imply that the RHS above is $f_{0\boldsymbol{\epsilon}}$ integrable. The conclusion of Lemma 8 follows from an application of DCT again. □

To prove Lemma 4, let $f_{\mathbf{U}|\mathbf{S}}$ denote the density of $\mathbf{U} = \mathbf{S}(\mathbf{X})\boldsymbol{\epsilon}$, where $\mathbf{S} = \mathrm{diag}(s_1, \dots, s_p)$. Then $f_{\mathbf{U}|\mathbf{X}} = f_{\mathbf{U}|\mathbf{S}(\mathbf{X})}$. We have $f_{\mathbf{U}|\mathbf{S}}(\mathbf{U}) = |\mathbf{S}|^{-1} f_{\boldsymbol{\epsilon}}(\mathbf{S}^{-1}\mathbf{U})$. This implies

$$\begin{aligned}\int f_{0\mathbf{U}|\mathbf{S}_0}(\mathbf{U}) \log \frac{f_{0\mathbf{U}|\mathbf{S}_0}(\mathbf{U})}{f_{\mathbf{U}|\mathbf{S}}(\mathbf{U})} d\mathbf{U} &= \int f_{0\mathbf{U}|\mathbf{S}_0}(\mathbf{U}) \log \frac{f_{0\mathbf{U}|\mathbf{S}_0}(\mathbf{U})}{f_{\mathbf{U}|\mathbf{S}_0}(\mathbf{U})} d\mathbf{U} + \int f_{0\mathbf{U}|\mathbf{S}_0}(\mathbf{U}) \log \frac{f_{\mathbf{U}|\mathbf{S}_0}(\mathbf{U})}{f_{\mathbf{U}|\mathbf{S}}(\mathbf{U})} d\mathbf{U} \\ &= \int f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \log \frac{f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})} d\boldsymbol{\epsilon} + \int f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \log \frac{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{|\mathbf{S}|^{-1}|\mathbf{S}_0| f_{\boldsymbol{\epsilon}}(\mathbf{S}^{-1}\mathbf{S}_0\boldsymbol{\epsilon})} d\boldsymbol{\epsilon}.\end{aligned}$$

Let $\delta > 0$ be given. By part 2 of Lemma 2, $\Pi_{\boldsymbol{\epsilon}}\{f_{\boldsymbol{\epsilon}} : d_{KL}(f_{0\boldsymbol{\epsilon}}, f_{\boldsymbol{\epsilon}}) < \delta/2\} > 0$. Let $\mathbf{s} = (s_1, \dots, s_p)^{\mathrm{T}}$ and $\mathbf{s}_0 = (s_{01}, \dots, s_{0p})^{\mathrm{T}}$. By Lemma 8, there exists $\eta > 0$ such that $\|\mathbf{s}_0 - \mathbf{s}\|_\infty < \eta$ implies $\int f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \log[f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) / \{|\mathbf{S}|^{-1}|\mathbf{S}_0| f_{\boldsymbol{\epsilon}}(\mathbf{S}^{-1}\mathbf{S}_0\boldsymbol{\epsilon})\}] \, d\boldsymbol{\epsilon} < \delta/2$ for every $f_{\boldsymbol{\epsilon}} \sim \Pi_{\boldsymbol{\epsilon}}$. Using a straightforward multivariate extension of Corollary 1, we have $\Pi_{\mathbf{V}}(||\mathbf{s}_0 - \mathbf{s}||_\infty < \eta) > 0$. Combining these results, $\Pi_{\mathbf{U}|\mathbf{V}}\{\sup_{\mathbf{X}\in\mathcal{X}} d_{KL}(f_{0\mathbf{U}|\mathbf{X}}, f_{\mathbf{U}|\mathbf{X}}) < \delta\} \ge \Pi_{\boldsymbol{\epsilon}}\{d_{KL}(f_{0\boldsymbol{\epsilon}}, f_{\boldsymbol{\epsilon}}) < \delta/2\} \, \Pi_{\mathbf{V}}(||\mathbf{s}_0 - \mathbf{s}||_\infty < \eta) > 0$. Hence the proof of part 2 of Lemma 4.

Part 1 of Lemma 4 follows trivially from part 2 of Lemma 4 since $||\mathbf{s}_0 - \mathbf{s}||_\infty < \eta$ implies $\|\mathbf{s}_0(\mathbf{X}) - \mathbf{s}(\mathbf{X})\|_\infty < \eta$ for any $\mathbf{X} \in \mathcal{X}$.

To prove part 3 of Lemma 4, note that

$$\begin{aligned} d_{KL}(f_{0,\mathbf{X},\mathbf{U}}, f_{\mathbf{X},\mathbf{U}}) &= \int_{\mathcal{X}\times\mathbb{R}^p} f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X}) f_{0,\mathbf{X}}(\mathbf{X}) \, \log \frac{f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X}) f_{0,\mathbf{X}}(\mathbf{X})}{f_{\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X})} d\mathbf{X} d\mathbf{U} \\ &= \int_{\mathcal{X}} f_{0,\mathbf{X}}(\mathbf{X}) \int_{\mathbb{R}^p} f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X}) \, \log \frac{f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X})}{f_{\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X})} d\mathbf{U} d\mathbf{X} + \int_{\mathcal{X}} f_{0,\mathbf{X}}(\mathbf{X}) \, \log \frac{f_{0,\mathbf{X}}(\mathbf{X})}{f_{\mathbf{X}}(\mathbf{X})} d\mathbf{X} \\ &\le \sup_{\mathbf{X}\in\mathcal{X}} d_{KL}\{f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X}), f_{\mathbf{U}|\mathbf{X}}(\mathbf{U}|\mathbf{X})\} + d_{KL}(f_{0\mathbf{X}}, f_{\mathbf{X}}). \end{aligned}$$

Part 3 of Lemma 4 now follows from part 2 of Lemma 4 and part 1 of Lemma 2.

S.7.4 Proof of Theorem 1

Let $d_H(f_0, f) = [\int \{f_0^{1/2}(\mathbf{Z}) - f^{1/2}(\mathbf{Z})\}^2 d\mathbf{Z}]^{1/2}$ denote the Hellinger distance between any two densities f_0 and f. From Chapter 1 of Ghosh and Ramamoorthi (2010), we have

$$d_H^2(f_0, f) \le ||f_0 - f||_1 \le 2 \, d_{KL}^{1/2}(f_0, f). \tag{S.8}$$

Using (S.8), we have,

$$||f_{0\mathbf{W}} - f_{\mathbf{W}}||_1 = \int |f_{0\mathbf{W}}(\mathbf{W}) - f_{\mathbf{W}}(\mathbf{W})| d\mathbf{W}$$

$$
\begin{aligned}
&= \int \left| \int f_{0\mathbf{X}}(\mathbf{X}) f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} - \int f_{\mathbf{X}}(\mathbf{X}) f_{\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} \right| d\mathbf{W} \\
&\leq \int \left| \int f_{0\mathbf{X}}(\mathbf{X}) f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} - \int f_{\mathbf{X}}(\mathbf{X}) f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} \right| d\mathbf{W} \\
&\quad + \int \left| \int f_{\mathbf{X}}(\mathbf{X}) f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} - \int f_{\mathbf{X}}(\mathbf{X}) f_{\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} \right| d\mathbf{W} \\
&\leq \int \int |f_{0\mathbf{X}}(\mathbf{X}) - f_{\mathbf{X}}(\mathbf{X})| f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) d\mathbf{X} d\mathbf{W} \\
&\quad + \int \int f_{\mathbf{X}}(\mathbf{X}) |f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) - f_{\mathbf{W}|\mathbf{X}}(\mathbf{W})| d\mathbf{X} d\mathbf{W} \\
&= \int |f_{0\mathbf{X}}(\mathbf{X}) - f_{\mathbf{X}}(\mathbf{X})| d\mathbf{X} + \int f_{\mathbf{X}}(\mathbf{X}) \int |f_{0\mathbf{W}|\mathbf{X}}(\mathbf{W}) - f_{\mathbf{W}|\mathbf{X}}(\mathbf{W})| d\mathbf{W} d\mathbf{X} \\
&= \int |f_{0\mathbf{X}}(\mathbf{X}) - f_{\mathbf{X}}(\mathbf{X})| d\mathbf{X} + \int f_{\mathbf{X}}(\mathbf{X}) \int |f_{0\mathbf{U}|\mathbf{X}}(\mathbf{W} - \mathbf{X}) - f_{\mathbf{U}|\mathbf{X}}(\mathbf{W} - \mathbf{X})| d\mathbf{W} d\mathbf{X} \\
&\leq ||f_{0\mathbf{X}} - f_{\mathbf{X}}||_1 + \sup_{\mathbf{X} \in \mathcal{X}} ||f_{0\mathbf{U}|\mathbf{X}} - f_{\mathbf{U}|\mathbf{X}}||_1 \\
&\leq 2\ d_{KL}^{1/2}(f_{0\mathbf{X}}, f_{\mathbf{X}}) + 2 \sup_{\mathbf{X} \in \mathcal{X}} d_{KL}^{1/2}(f_{0\mathbf{U}|\mathbf{X}}, f_{\mathbf{U}|\mathbf{X}}).
\end{aligned}
$$

The proof of Theorem 1 follows by combining part 1 of Lemma 2 and part 2 of Lemma 4.

S.8 Additional Figures

We first present, in Subsection S.8.1, some additional figures summarizing the results of the simulation experiments for diagonal covariance matrices discussed in Section 6 of the main paper. Then in Subsection S.8.1, we present figures that summarize the results of simulation experiments for covariance matrices with AR structure. Finally in Subsection S.8.3, we present some additional figures summarizing the results of the EATS data set analyzed in Section 7 of the main paper.

S.8.1 Additional Figures Summarizing the Results of the Simulation Experiments for Diagonal Covariance Structure

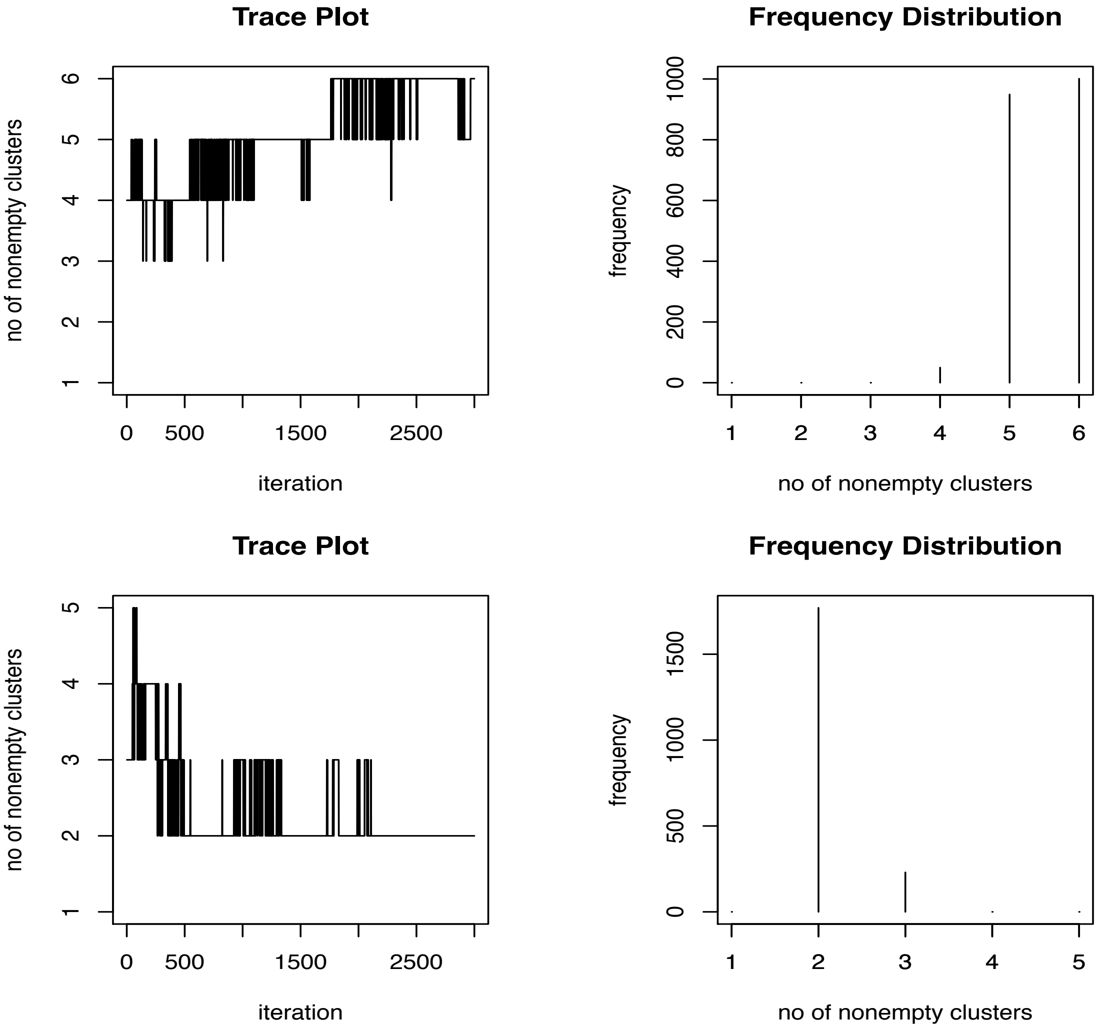


Figure S.5: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 6 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure 5, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

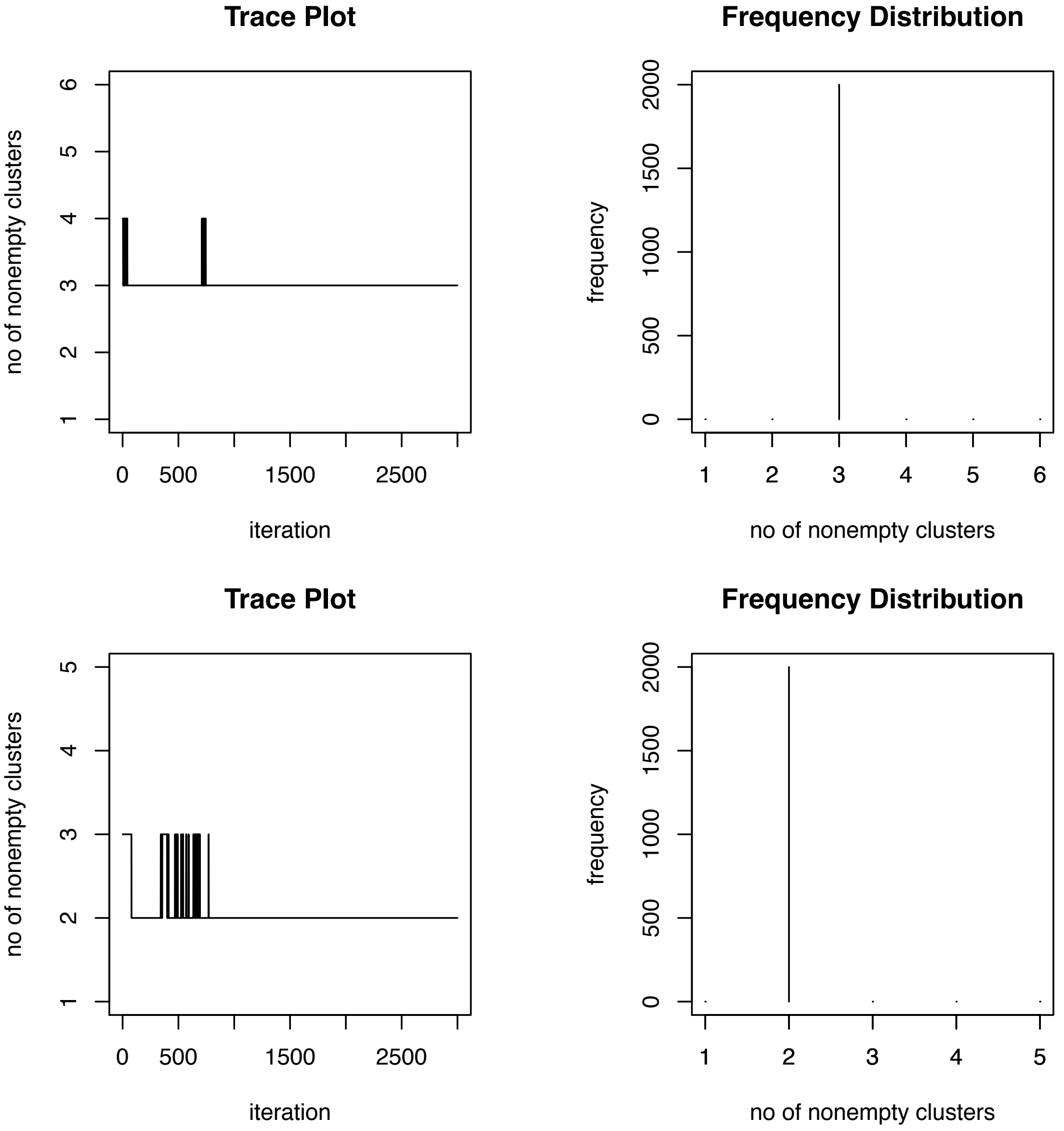


Figure S.6: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 6 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure 6, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

S.8.2 Additional Figures Summarizing the Results of the Simulation Experiments for AR Covariance Structure

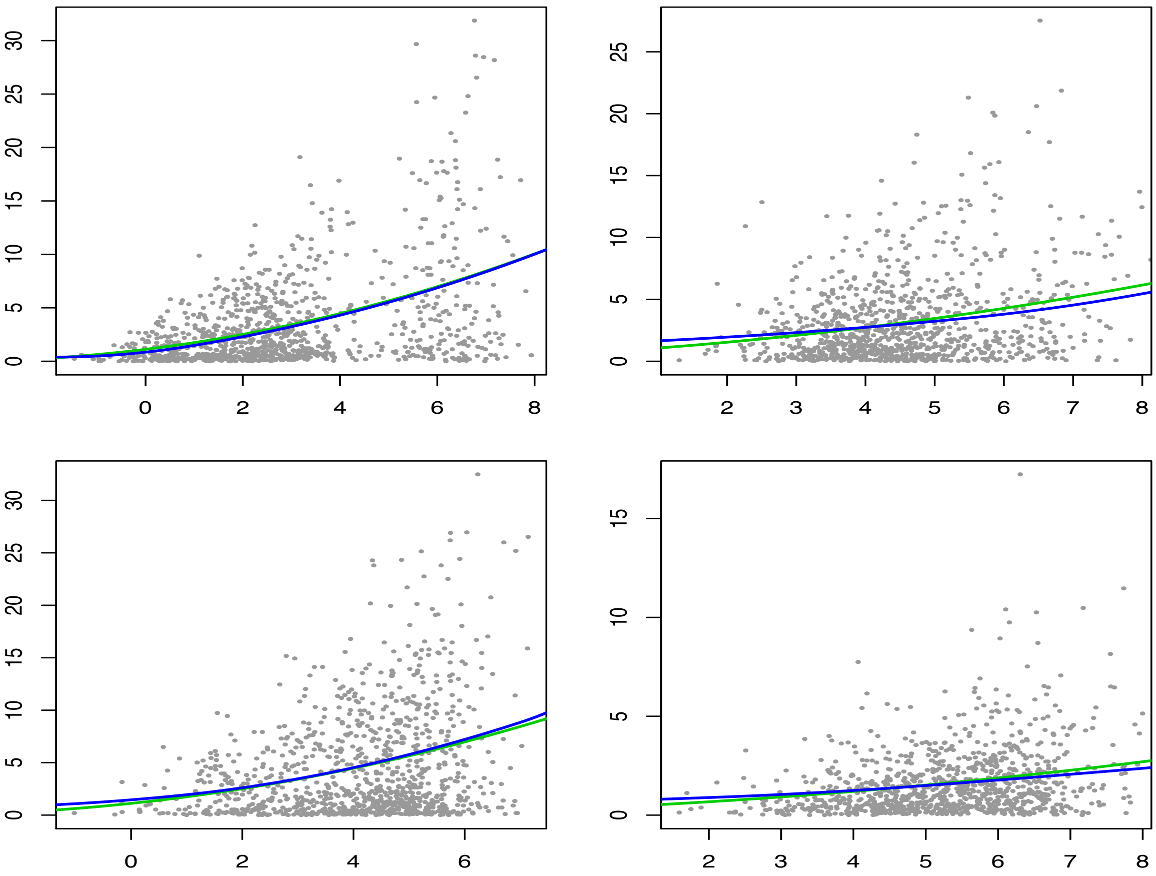


Figure S.7: Results for the variance functions $s^2(X)$ produced by the univariate density deconvolution method for each component of $\mathbf{X}$ for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets for the MIW (mixtures with inverse Wishart priors) method. For each component of $\mathbf{X}$, the true variance function is $s^2(X) = (1 + X/4)^2$. See Section 2.2.2 and Section S.3 for additional details. In each panel, the true (lighter shaded green lines) and the estimated (darker shaded blue lines) variance functions are superimposed over a plot of subject specific sample means vs subject specific sample variances. The figure is in color in the electronic version of this article.

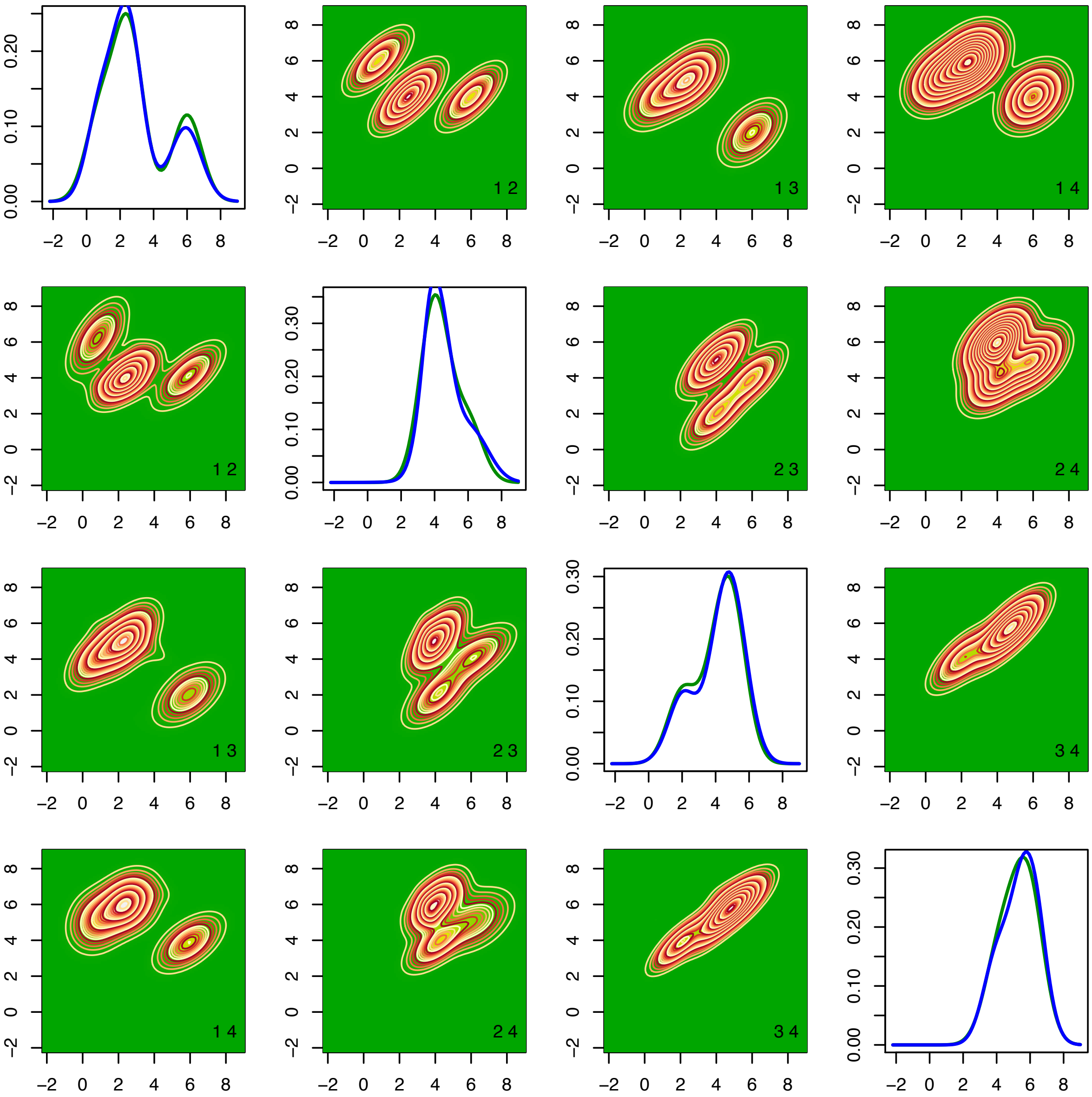


Figure S.8: Results for the $f_{\mathbf{X}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities f_{X_i,X_j} are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

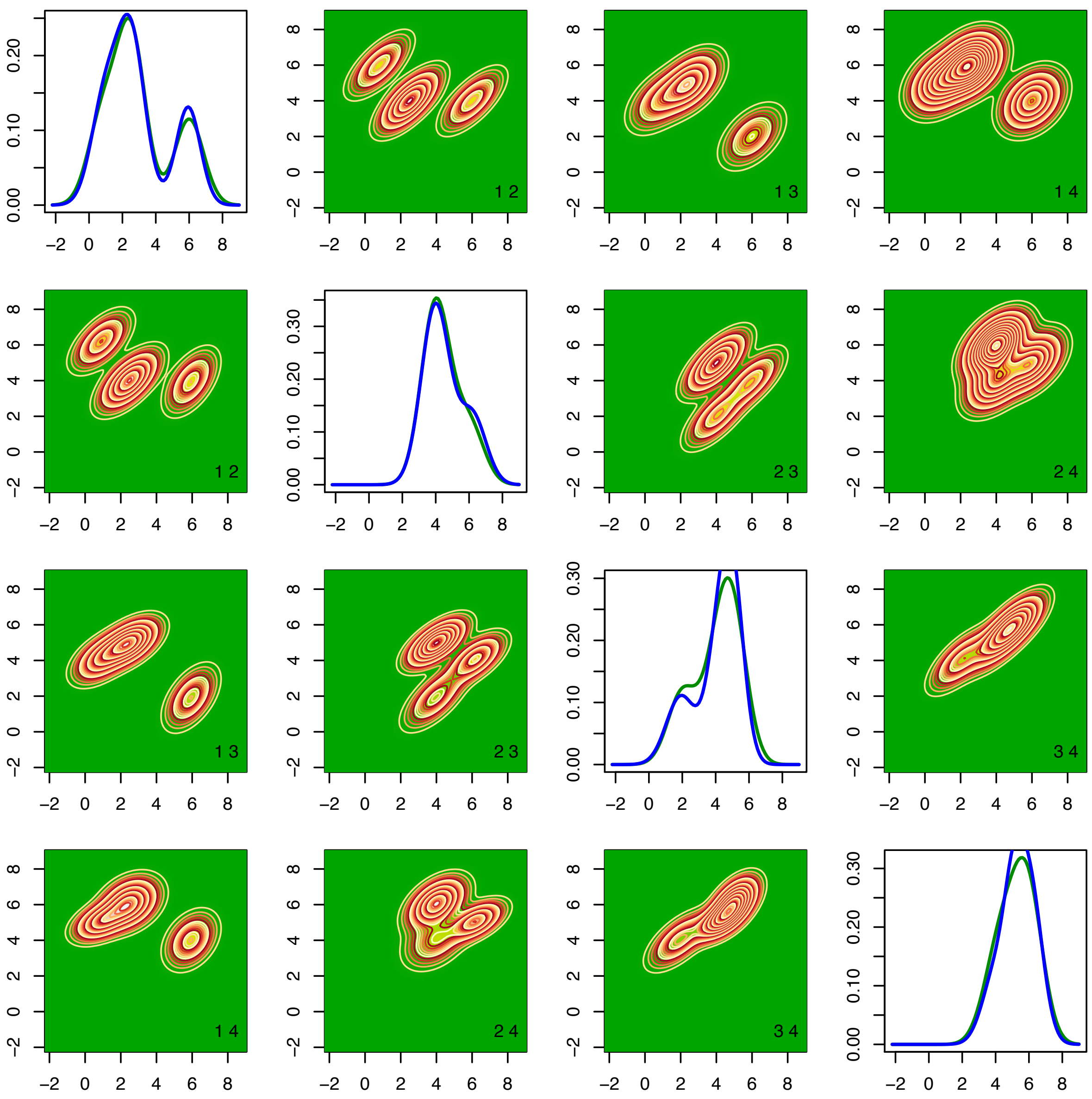


Figure S.9: Results for the $f_{\mathbf{X}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities f_{X_i,X_j} are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

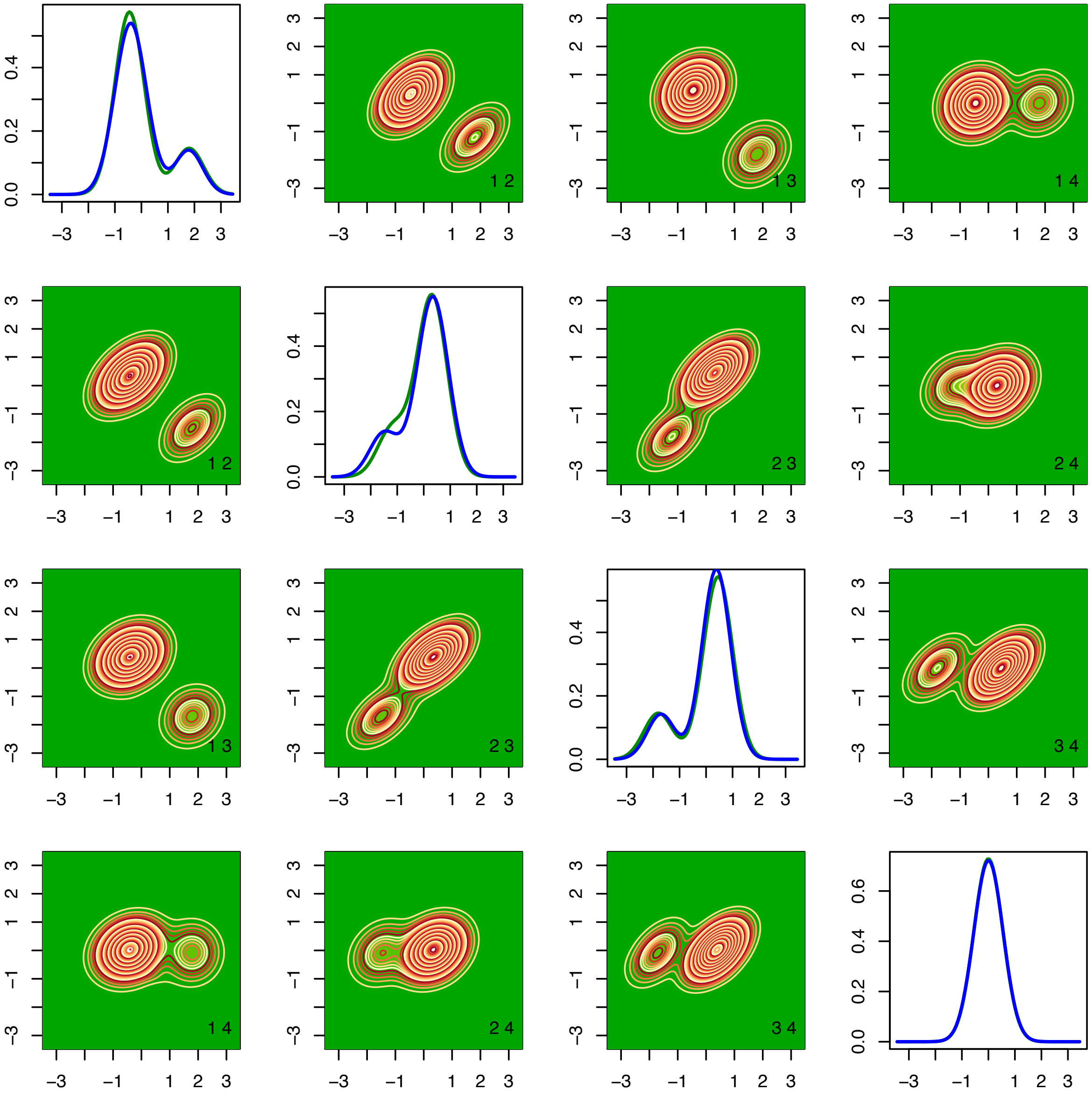


Figure S.10: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities $f_{\epsilon_i,\epsilon_j}$ are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

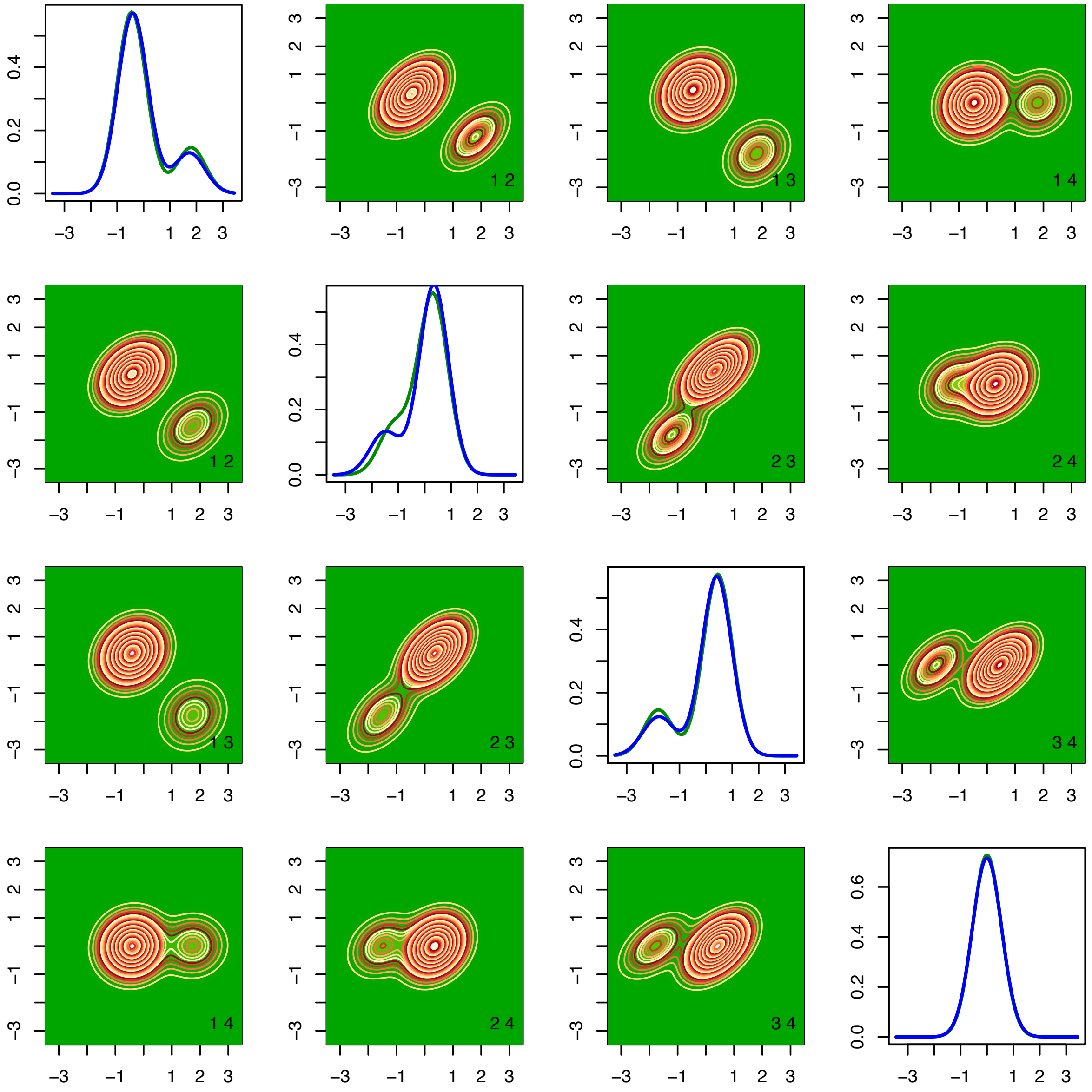


Figure S.11: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities $f_{\epsilon_i,\epsilon_j}$ are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

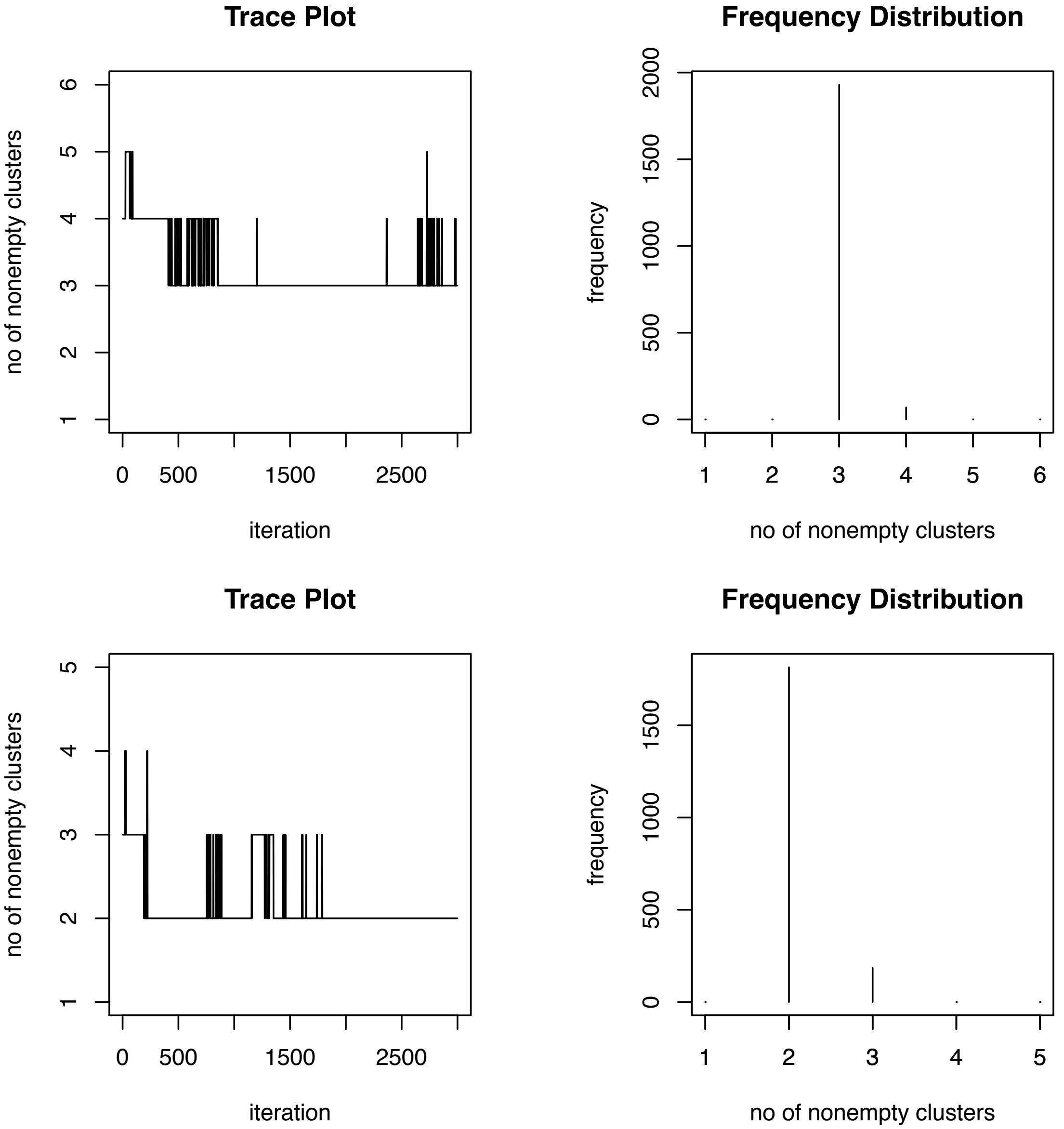


Figure S.12: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). See Section 6 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure S.10, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

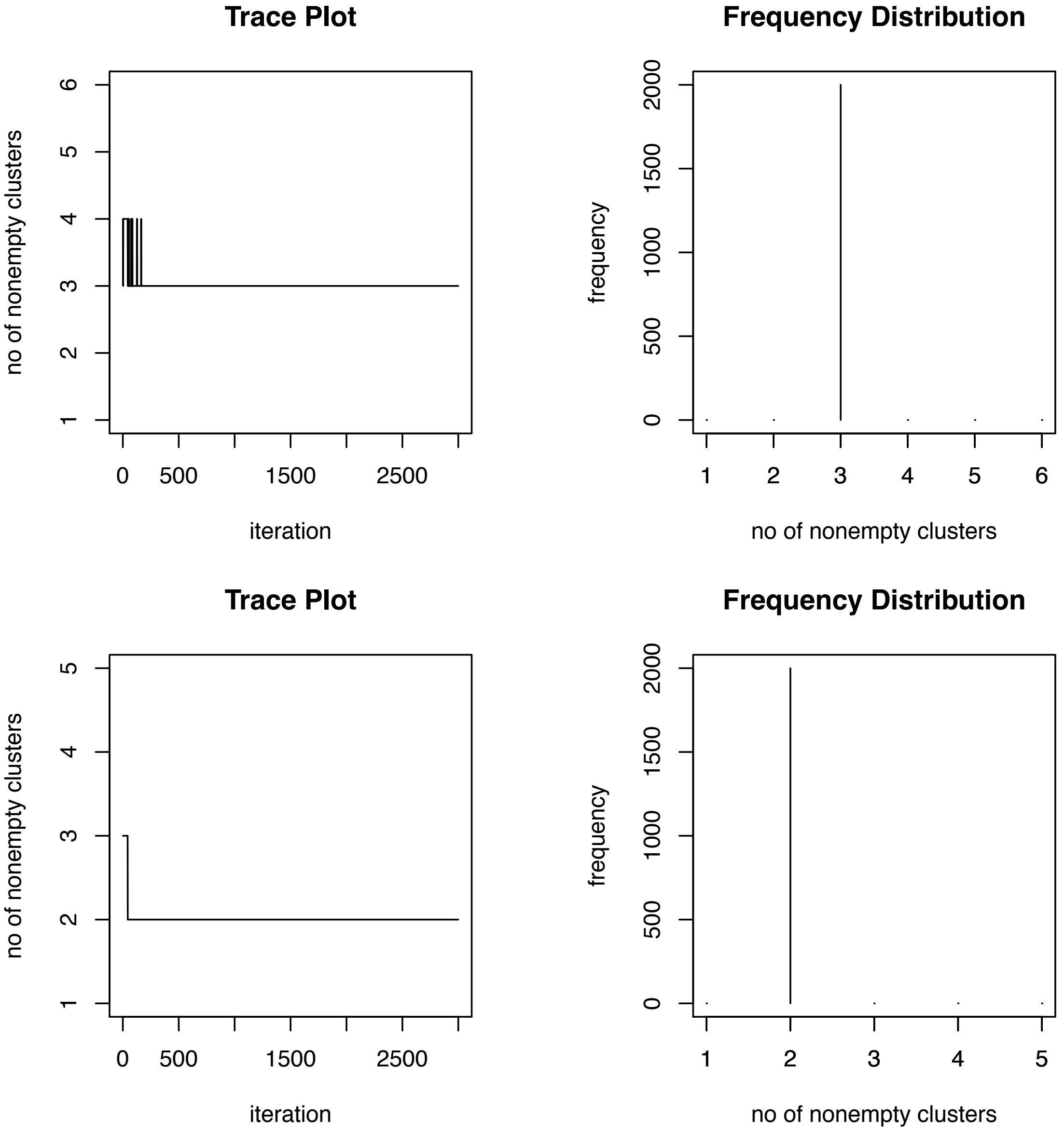
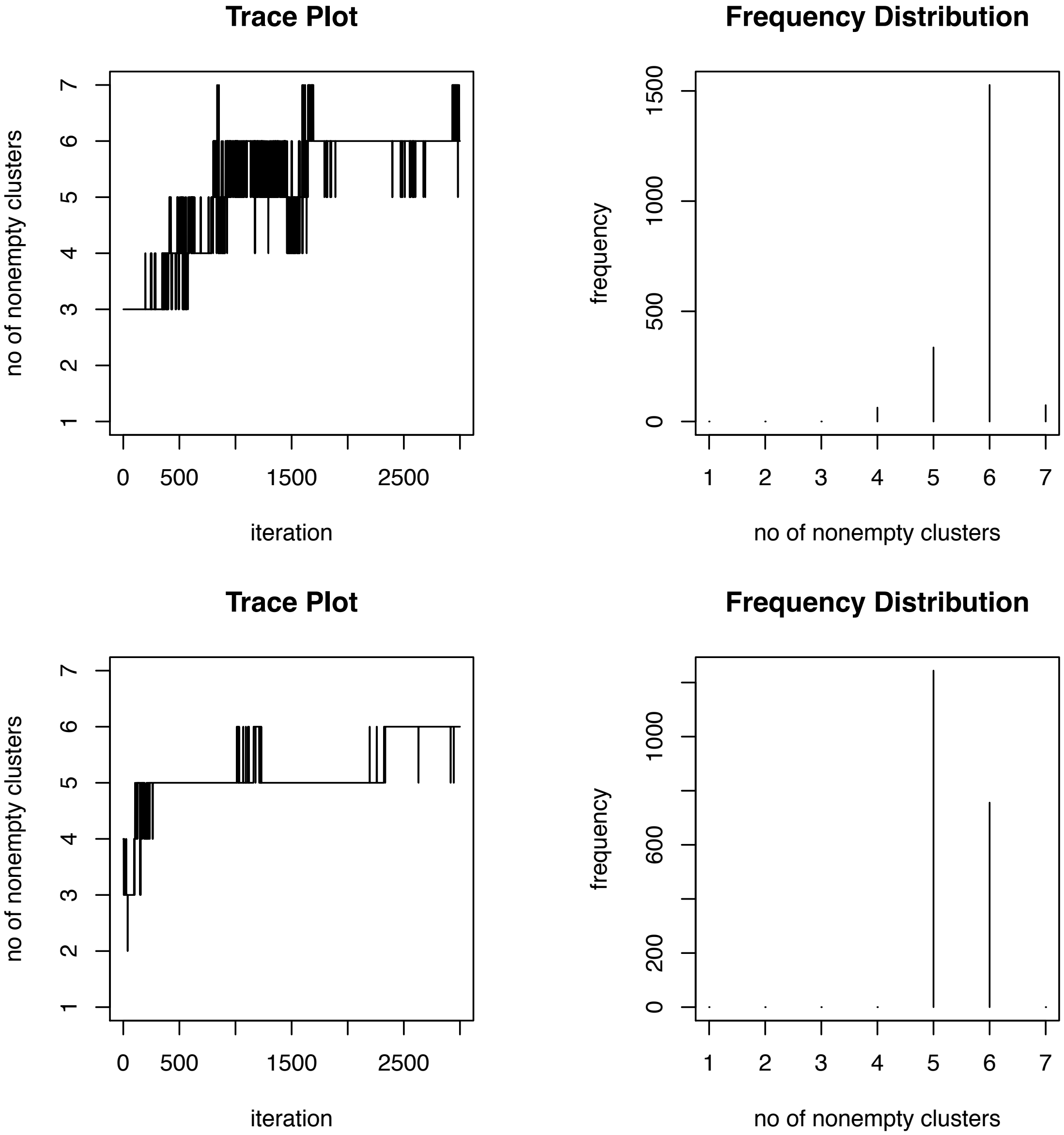


Figure S.13: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f_{\boldsymbol{\epsilon}}^{(2)}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). See Section 6 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure S.11, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

S.8.3 Additional Figures Summarizing the Results for the EATS Data Set Analyzed in Section 7 of the Main Paper


Trace Plot

Frequency Distribution

Trace Plot

Frequency Distribution

Figure S.14: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the EATS data example. See Section 7 for additional details. The number of mixture components for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = K_{\boldsymbol{\epsilon}} = 7$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$.

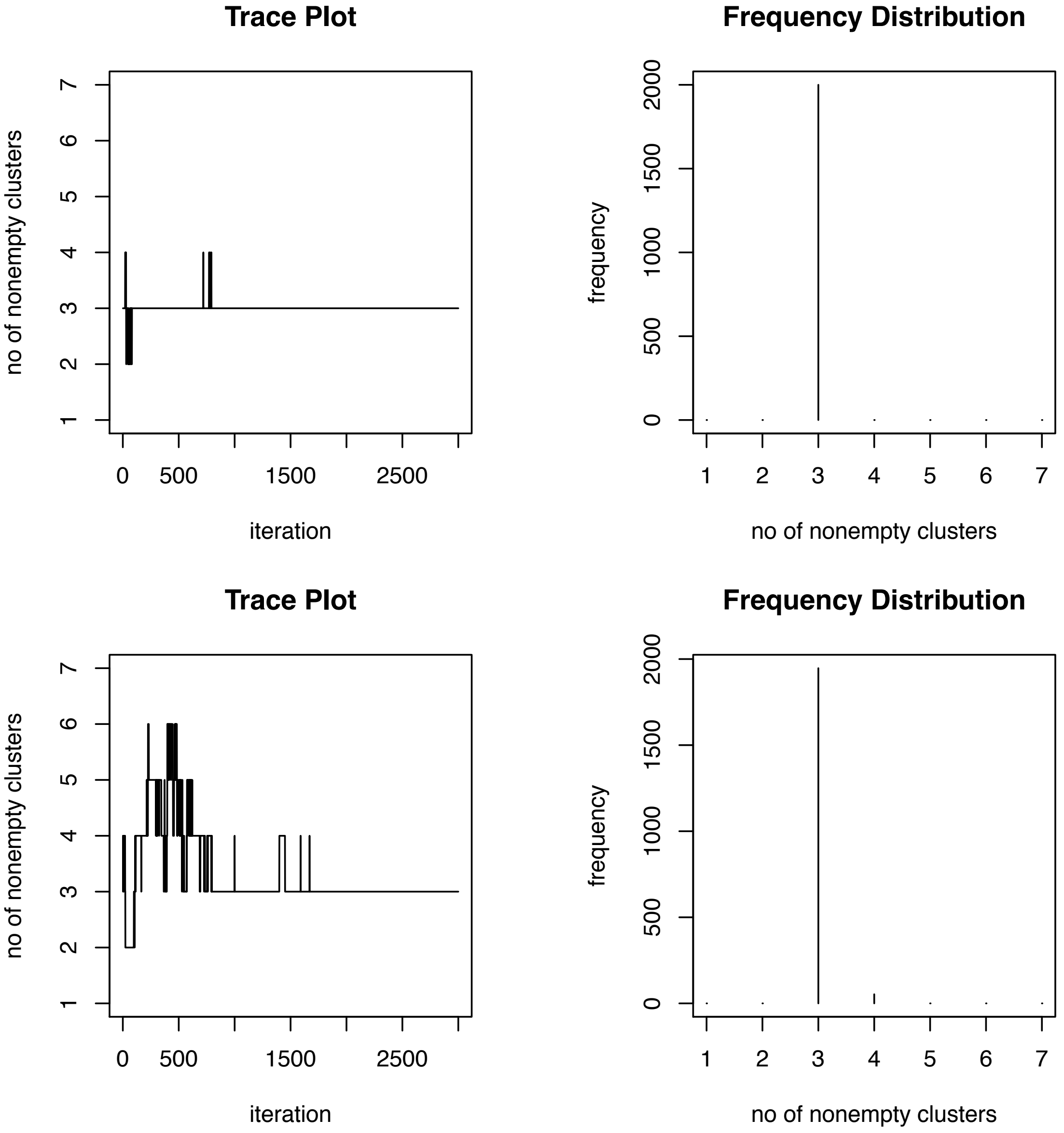


Figure S.15: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the EATS data example. See Section 7 for additional details. The number of mixture components for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = K_{\boldsymbol{\epsilon}} = 7$. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$.

S.9 Additional Simulation Experiments

This section presents the results of additional simulation experiments for multivariate t and multivariate Laplace distributed measurement errors. Cases when $f_{\mathbf{X}}$ is multivariate t or mixture of multivariate t are also considered. For easy reference, brief descriptions of these distributions are provided below.

S.9.1 Multivariate t Distribution

A random variable Z following a Student's t-distribution with degrees of freedom ν, mean μ and variance $\nu b/(\nu-2)$ can be represented as $Z = \mu + \nu^{1/2}b^{1/2}X/Y^{1/2}$, where Y and X are independent, Y follows a chi-square distribution with ν degrees of freedom, denoted by $Y \sim \chi^2_\nu$, and X follows a standard normal distribution. A natural extension to multivariate set up is given by $\mathbf{Z} = \boldsymbol{\mu} + \nu^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{X}/Y^{1/2}$, where $Y \sim \chi^2_\nu$ and $\mathbf{X} \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I})$ independently. The random vector $\mathbf{Z}$ is then said to follow a multivariate t-distribution (Kotz and Nadarajah, 2004) with degrees of freedom ν, mean $\boldsymbol{\mu}$ and covariance $\nu\boldsymbol{\Sigma}/(\nu-2)$, denoted by $\mathrm{MVT}_p(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The above characterization can be used to sample from a $\mathrm{MVT}_p(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ density. The density of $\mathbf{Z}$ is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{\Gamma\{(\nu+p)/2\}}{\Gamma(\nu/2)(\nu\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \cdot \{1 + (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})/\nu\}^{-(\nu+p)/2}.$$

The characteristic function is given by

$$\phi(\mathbf{t}) = \exp(i\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}) \cdot \frac{||\nu^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{t}||^{\nu/2}}{2^{\nu/2-1}\Gamma(\nu/2)} \cdot H_{\nu/2}(||\nu^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{t}||), \quad \mathbf{t} \in \mathbb{R}^p,$$

where H_α denotes a McDonald's function of order $\alpha(> 1/2)$ and admits the integral representation

$$H_\alpha(t) = (2/t)^\alpha \cdot \frac{\Gamma(\alpha+1/2)}{\pi^{1/2}} \int_0^\infty (1+u^2)cos(tu)du, \quad t > 0.$$

When $\boldsymbol{\Sigma} = \mathbf{I}$, the identity matrix, the components Z_i and Z_j are uncorrelated, but not statistically independent. With $\boldsymbol{\mu} = (\mu_1 \ldots, \mu_p)^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = ((\sigma_{ij}))$, the i^{th} random variable Z_i marginally follows a univariate Student's t-distribution with degrees of freedom ν, mean μ_i and variance $\nu\sigma_{ii}/(\nu-2)$.

S.9.2 Multivariate Laplace Distribution

A random variable Z following a Laplace distribution with mean μ and variance b has the density

$$f_Z(z) = (2b)^{-1/2}\exp(-2^{1/2}b^{-1/2}\left|z-\mu\right|).$$

Z can be represented as $Z = \mu + Y^{1/2}b^{1/2}X$, where Y and X are independent and follow standard exponential and standard normal distributions, respectively. A natural extension to multivariate set up is given by $\mathbf{Z} = \boldsymbol{\mu} + Y^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{X}$, where Y follows a standard exponential density and $\mathbf{X} \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I})$ independently of Y. The random vector $\mathbf{Z}$ is then said to follow a multivariate Laplace distribution (Eltoft, et al. 2006) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, denoted by $\mathrm{MVL}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The above characterization can be used to sample from a $\mathrm{MVL}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density. The density of $\mathbf{Z}$ is then given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{2}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \cdot \frac{K_{p/2-1}\{2^{1/2}h^{1/2}(\mathbf{z})\}}{\{h(\mathbf{z})/2\}^{p/4-1/2}},$$

where $h(\mathbf{z}) = (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})$ and K_m denotes modified Bessel functions of the second kind of order m. Using asymptotic formula for the Bessel functions, namely $K_m(z) = \{\pi/(2z)\}^{1/2}\exp(-z)$ as $|z| \to \infty$, we have

$$f_{\mathbf{Z}}(\mathbf{z}) \approx \frac{\pi^{1/2}}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \cdot \frac{2^{(p-1)/4}}{h^{(p-1)/4}(\mathbf{z})} \cdot \exp\{-2^{1/2}h^{1/2}(\mathbf{z})\}.$$

The characteristic function is given by $\phi(\mathbf{t}) = \exp(i\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu})(1+\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}/2)^{-1}$ for $\mathbf{t} \in \mathbb{R}^p$. For $p > 1$, the density has a singularity at $\boldsymbol{\mu}$. When $\boldsymbol{\Sigma} = \mathbf{I}$, the identity matrix, the components Z_i and Z_j are uncorrelated, but not statistically independent. With $\boldsymbol{\mu} = (\mu_1 \ldots, \mu_p)^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = ((\sigma_{ij}))$, the i^{th} random variable Z_i marginally follows a univariate Laplace distribution with mean μ_i and variance σ_{ii}.

S.9.3 Summary of Results

The results of the simulation experiments the measurement errors are distributed according to $f_{\boldsymbol{\epsilon}}^{(3)} = \mathrm{MVT}_4(6, \mathbf{0}, \boldsymbol{\Sigma})$ and $f_{\boldsymbol{\epsilon}}^{(4)} = \mathrm{MVL}_4(\mathbf{0}, \boldsymbol{\Sigma})$ probability laws independently of $\mathbf{X}$ are presented in Table S.1. The results for conditionally heteroscedastic measurement errors are presented in Table S.2. In both cases, $\mathbf{X}$ is distributed according to the mixture of multivariate normals described in Section 6 of the main paper. As in the main paper, in each case four different choices for the covariance matrix $\boldsymbol{\Sigma}$ were considered. The general patterns of the estimated MISEs are similar to that observed in Table 2 of the main paper where the true measurement error distributions were finite mixtures of multivariate normal kernels. While in theory the MLFA model described in the main paper can approximate distributions like the multivariate Laplace that puts significant mass around the origin, in practice, since it assumes $\boldsymbol{\Omega}_k = \boldsymbol{\Omega} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$ for all k, it often smooths out the spikes at the origin. A mild variation, referred to as the MLFA_2 model, that instead assumes $\boldsymbol{\Omega}_k = \sigma_k^2\mathrm{I}_p$ and results in slight improvement in the MISE performance is also included in Table S.1 and Table S.2. For the simulation experiments and the real data analysis presented in the main text, the two versions of the MLFA model perform very similarly and the latter version was not included. Results for conditionally heteroscedastic multivariate Laplace errors with diagonal covariance structure are summarized in Figures S.16-S.22 with observations similar to those discussed in Section 6 of the main paper.

True Error Distribution	Covariance Structure	Sample Size	MISE $\times 10^4$			
			MLFA$_2$	MLFA	MIW	Naive
(c) Multivariate t	I	500	**1.06**	1.38	3.98	12.32
		1000	**0.53**	0.65	1.54	9.91
	LF	500	**6.62**	8.26	7.57	47.22
		1000	4.73	5.78	**3.65**	45.70
	AR	500	12.69	13.56	**6.14**	40.76
		1000	11.36	9.16	**3.45**	39.59
	EXP	500	7.84	8.42	**5.00**	26.85
		1000	6.26	6.64	**2.38**	26.04
(d) Multivariate Laplace	I	500	**1.08**	1.32	3.08	8.22
		1000	**0.50**	0.63	1.20	6.25
	LF	500	**4.41**	5.57	5.66	32.31
		1000	**2.38**	3.53	2.84	31.10
	AR	500	8.38	8.72	**5.14**	27.30
		1000	6.08	6.19	**2.56**	26.19
	EXP	500	5.24	5.67	**4.14**	17.57
		1000	3.58	4.17	**1.98**	16.86

Table S.1: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models for **homoscedastic** errors compared with a naive method that ignores measurement errors for different measurement error distributions. See Section 2 and Section S.9 for additional details. The minimum value in each row is highlighted.

True Error Distribution	Covariance Structure	Sample Size	MISE $\times 10^4$			
			MLFA$_2$	MLFA	MIW	Naive
(c) Multivariate t	I	500	**2.78**	3.25	24.48	19.10
		1000	**1.39**	1.53	13.40	17.75
	LF	500	**12.65**	14.72	52.77	69.64
		1000	**6.71**	8.43	25.66	66.49
	AR	500	**20.54**	23.2	43.22	64.07
		1000	**13.53**	18.41	21.42	59.81
	EXP	500	**11.56**	14.12	37.68	43.57
		1000	**8.19**	11.97	18.22	41.66
(d) Multivariate Laplace	I	500	**1.81**	2.32	9.60	10.31
		1000	**0.97**	1.20	4.20	8.86
	LF	500	**7.33**	10.30	17.52	41.89
		1000	**3.99**	5.28	7.65	40.93
	AR	500	**9.79**	14.13	15.64	35.50
		1000	**5.54**	9.32	6.59	34.91
	EXP	500	**7.26**	9.90	13.93	23.71
		1000	**3.90**	5.12	5.19	22.78

Table S.2: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models for **conditionally heteroscedastic** errors compared with a naive method that ignores measurement errors for different measurement error distributions. See Section 2 and Section S.9 for additional details. The minimum value in each row is highlighted.

We also extend the simulation experiments to scenarios when $\mathbf{X}$ is distributed according to (B) $f_{\mathbf{X}}^{(3)} = \mathrm{MVT}_4(6, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}}), \boldsymbol{\mu}_{\mathbf{X}} = (2,2,2,2)^{\mathrm{T}}$, (C) $f_{\mathbf{X}}^{(4)} = \sum_{k=1}^{2} \pi_{\mathbf{X},k} \mathrm{MVT}_4(6, \boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X}})$, $\boldsymbol{\pi}_{\mathbf{X}} = (0.75, 0.25)^{\mathrm{T}}, \boldsymbol{\mu}_{\mathbf{X},1} = (2,4,2,2)^{\mathrm{T}}, \boldsymbol{\mu}_{\mathbf{X},2} = (4,2,4,2)^{\mathrm{T}}$. In each case, four different choices for $\boldsymbol{\Sigma}_{\mathbf{X}}$ are considered as in Section 6 of the main paper. We focus on the case when the measurement errors are conditionally heteroscedastic. Results are presented in Tables S.3 and S.4.

True Distribution of Interest $f_{\mathbf{X}}$	True Error Distribution $f_{\boldsymbol{\epsilon}}$	Covariance Structure	Sample Size	MISE $\times 10^4$		
				MLFA$_2$	MIW	Naive
(B) Multivariate t	(a) Multivariate Normal	I	500	**4.35**	20.36	18.17
			1000	**2.36**	13.14	12.65
		LF	500	**21.31**	78.22	75.42
			1000	**15.57**	52.73	67.77
		AR	500	**33.18**	59.77	63.33
			1000	**29.29**	51.11	53.40
		EXP	500	**19.58**	40.72	44.83
			1000	**17.78**	32.01	37.58
	(b) Mixture of Multivariate Normals	I	500	**5.16**	27.21	38.03
			1000	**2.87**	18.17	35.99
		LF	500	**27.89**	73.75	159.29
			1000	**19.27**	53.66	161.77
		AR	500	**38.41**	81.77	159.34
			1000	**34.22**	55.25	156.05
		EXP	500	**21.95**	45.76	100.33
			1000	**18.14**	37.72	99.09
	(c) Multivariate t	I	500	**4.16**	27.73	23.42
			1000	**2.34**	19.87	20.36
		LF	500	**22.83**	91.04	90.39
			1000	**14.03**	85.33	89.31
		AR	500	**40.60**	76.40	86.87
			1000	**36.93**	70.76	75.19
		EXP	500	**26.36**	55.65	61.25
			1000	**18.51**	40.46	49.52
	(d) Multivariate Laplace	I	500	**3.93**	16.48	16.14
			1000	**1.81**	6.85	14.02
		LF	500	**16.36**	47.19	70.22
			1000	**12.13**	27.64	59.48
		AR	500	**29.46**	42.44	63.79
			1000	**18.81**	21.19	47.92
		EXP	500	**19.00**	34.74	39.64
			1000	**13.30**	16.24	32.76

Table S.3: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models for **conditionally heteroscedastic** errors compared with a naive method that ignores measurement errors for different measurement error distributions. See Section 2 and Section S.9 for additional details. The minimum value in each row is highlighted.

True Distribution of Interest $f_{\mathbf{X}}$	True Error Distribution $f_{\boldsymbol{\epsilon}}$	Covariance Structure	Sample Size	MISE $\times 10^4$		
				MLFA$_2$	MIW	Naive
(C) Mixture of Multivariate t	(a) Multivariate Normal	I	500	**4.84**	13.68	12.43
			1000	**2.82**	7.41	10.15
		LF	500	**21.62**	30.01	47.95
			1000	**13.40**	19.72	44.97
		AR	500	**22.56**	29.35	43.99
			1000	**19.80**	25.59	39.63
		EXP	500	**18.36**	27.27	28.00
			1000	**13.41**	17.73	25.14
	(b) Mixture of Multivariate Normals	I	500	**5.39**	14.64	22.90
			1000	**2.80**	10.77	21.55
		LF	500	**24.48**	32.87	98.00
			1000	**15.62**	20.52	98.79
		AR	500	**26.73**	31.09	90.78
			1000	**23.44**	29.06	91.24
		EXP	500	**19.56**	25.39	58.83
			1000	**13.90**	18.29	59.93
	(c) Multivariate t	I	500	**4.91**	18.09	16.30
			1000	**2.89**	11.59	14.00
		LF	500	**23.50**	33.79	60.18
			1000	**15.85**	25.83	58.20
		AR	500	**26.98**	33.78	54.07
			1000	**22.04**	29.77	51.64
		EXP	500	**18.62**	24.00	36.26
			1000	**12.64**	18.57	33.61
	(d) Multivariate Laplace	I	500	**4.76**	9.34	15.96
			1000	**2.33**	5.04	13.96
		LF	500	**16.59**	22.54	65.33
			1000	**11.69**	13.41	59.25
		AR	500	**24.73**	26.21	58.87
			1000	**15.71**	17.48	47.62
		EXP	500	**14.26**	19.12	34.53
			1000	**10.96**	13.25	32.47

Table S.4: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models for **conditionally heteroscedastic** errors compared with a naive method that ignores measurement errors for different measurement error distributions. See Section 2 and Section S.9 for additional details. The minimum value in each row is highlighted.

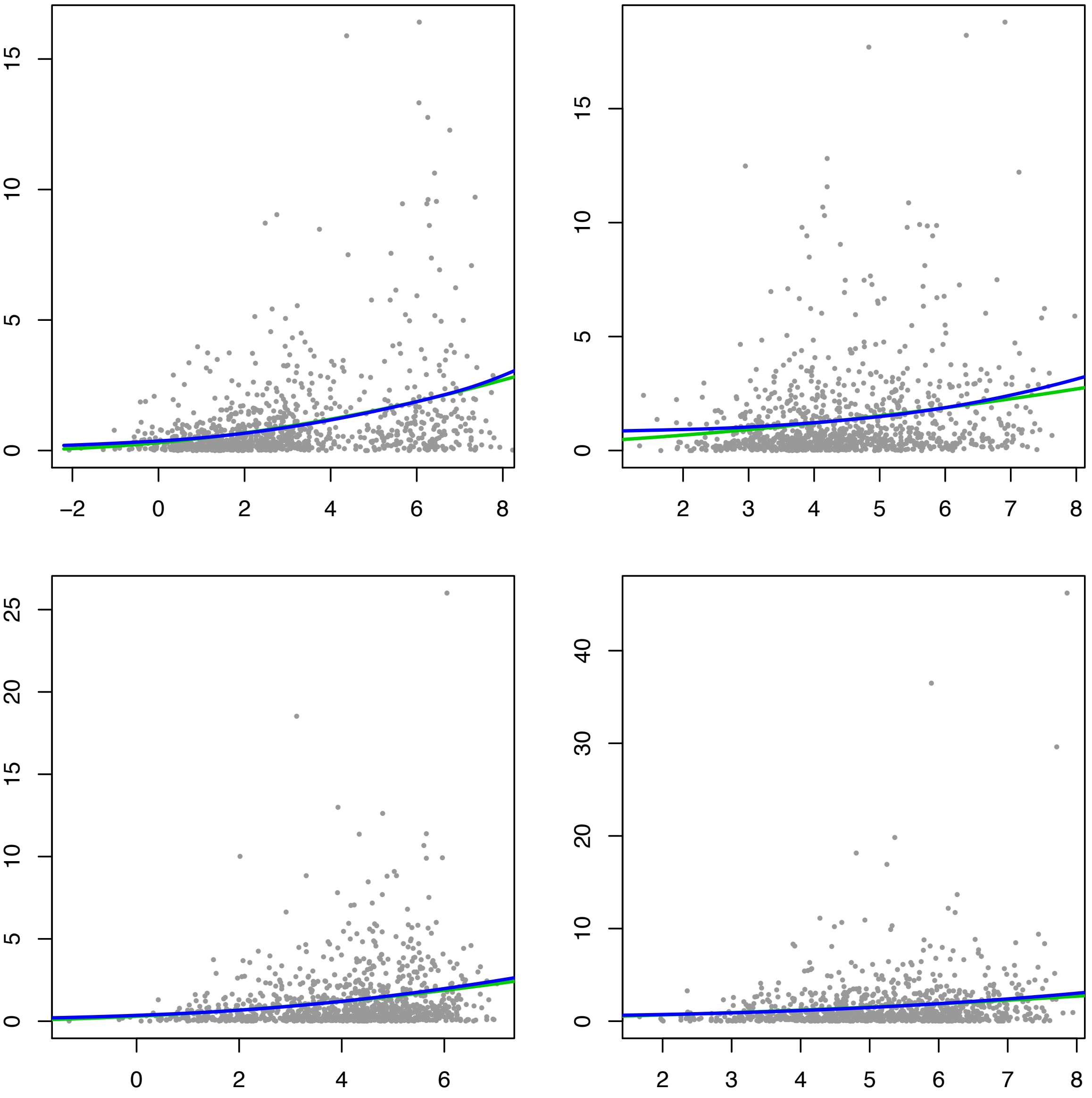


Figure S.16: Results for the variance functions $s^2(X)$ produced by the univariate density deconvolution method for each component of $\mathbf{X}$ for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets for the MIW (mixtures with inverse Wishart priors) method. For each component of $\mathbf{X}$, the true variance function is $s^2(X) = (1 + X/4)^2$. See Section 2.2.2 and Section S.3 for additional details. In each panel, the true (lighter shaded green lines) and the estimated (darker shaded blue lines) variance functions are superimposed over a plot of subject specific sample means vs subject specific sample variances. The figure is in color in the electronic version of this article.

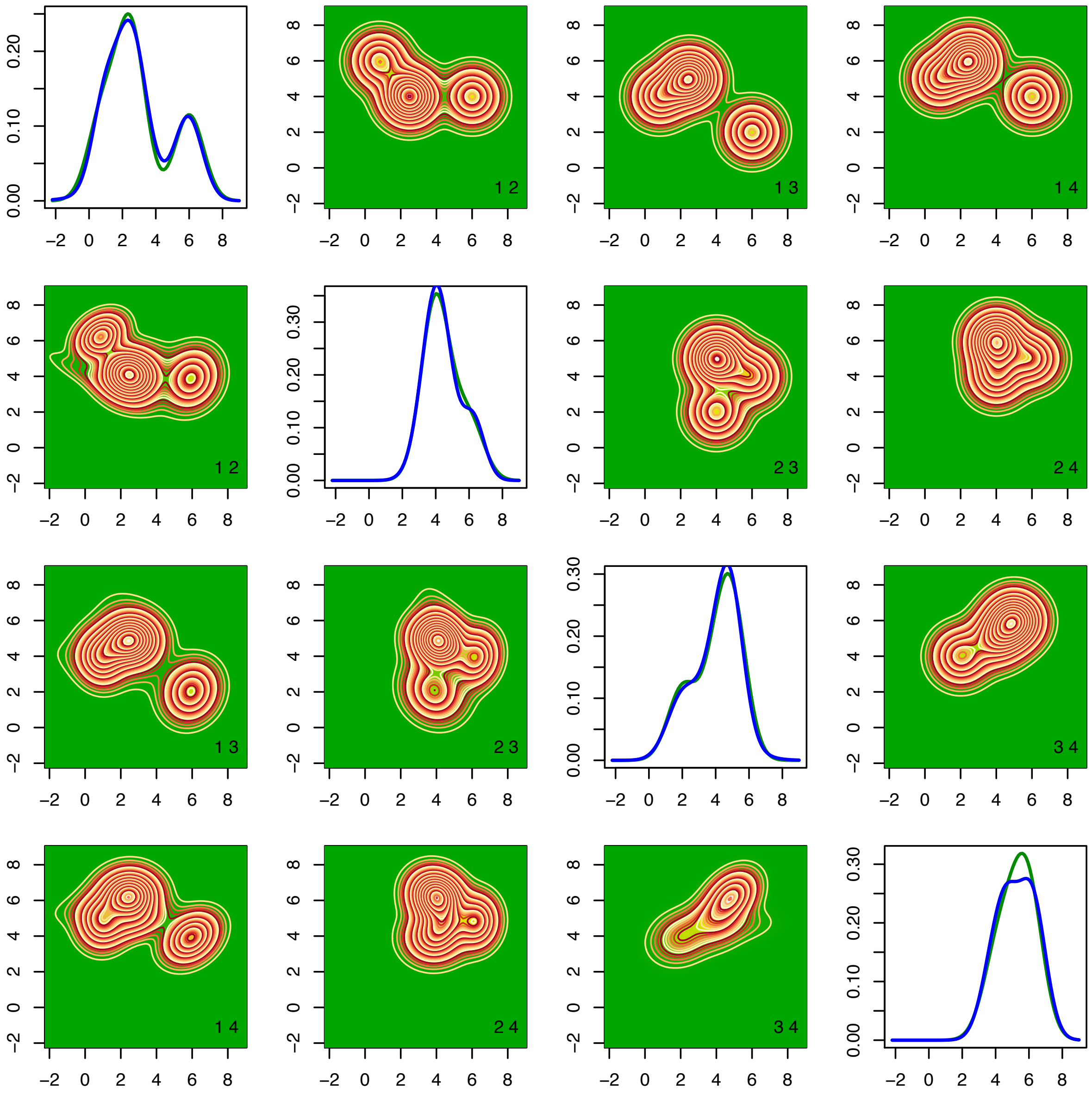


Figure S.17: Results for the $f_{\mathbf{X}}$ produced by the MIW (mixtures with inverse Wishart priors) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 and Section S.9 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities f_{X_i,X_j} are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

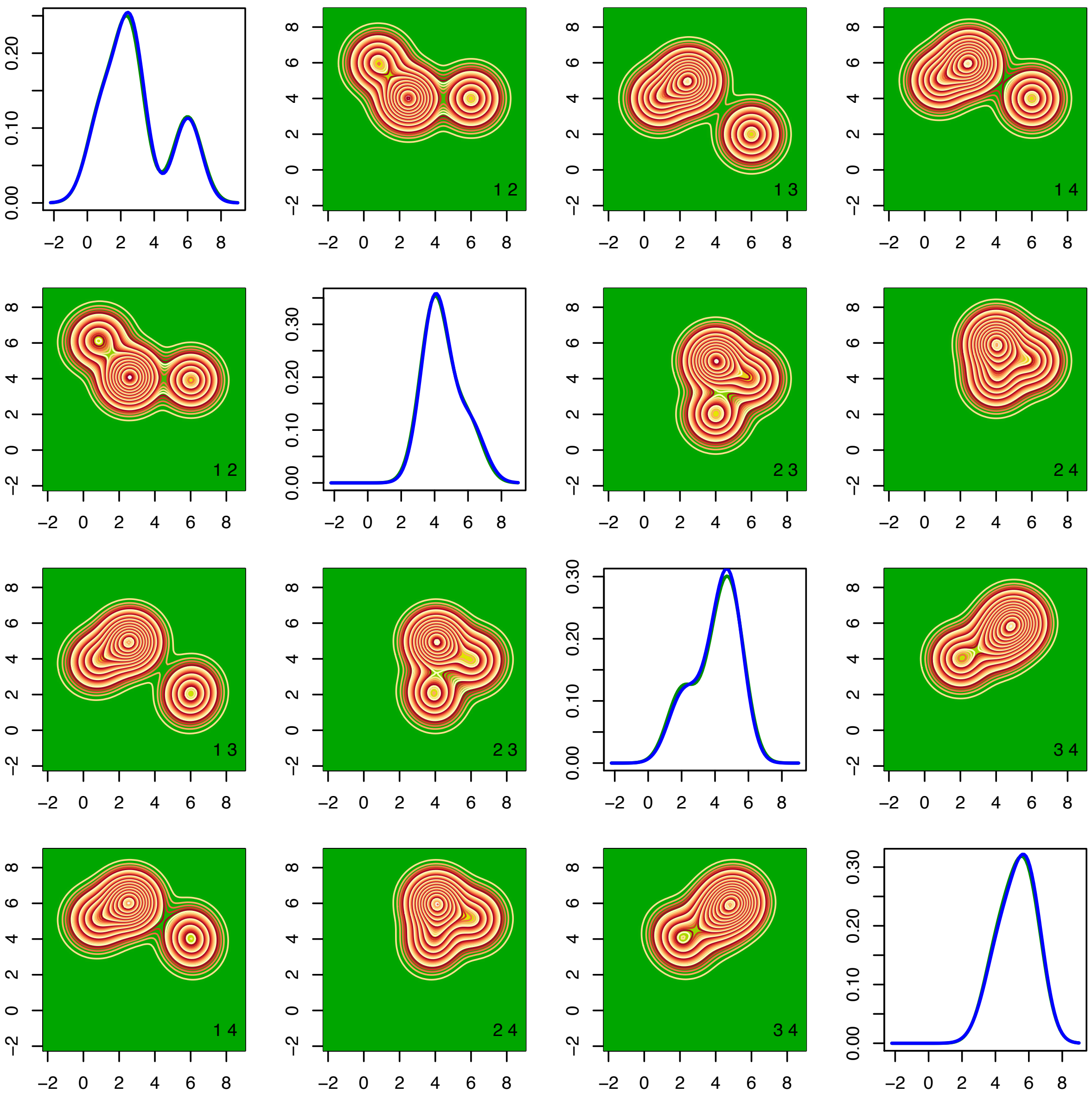


Figure S.18: Results for the $f_{\mathbf{X}}$ produced by the MLFA$_2$ (mixtures of latent factor analyzers) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 and Section S.9 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities f_{X_i,X_j} are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

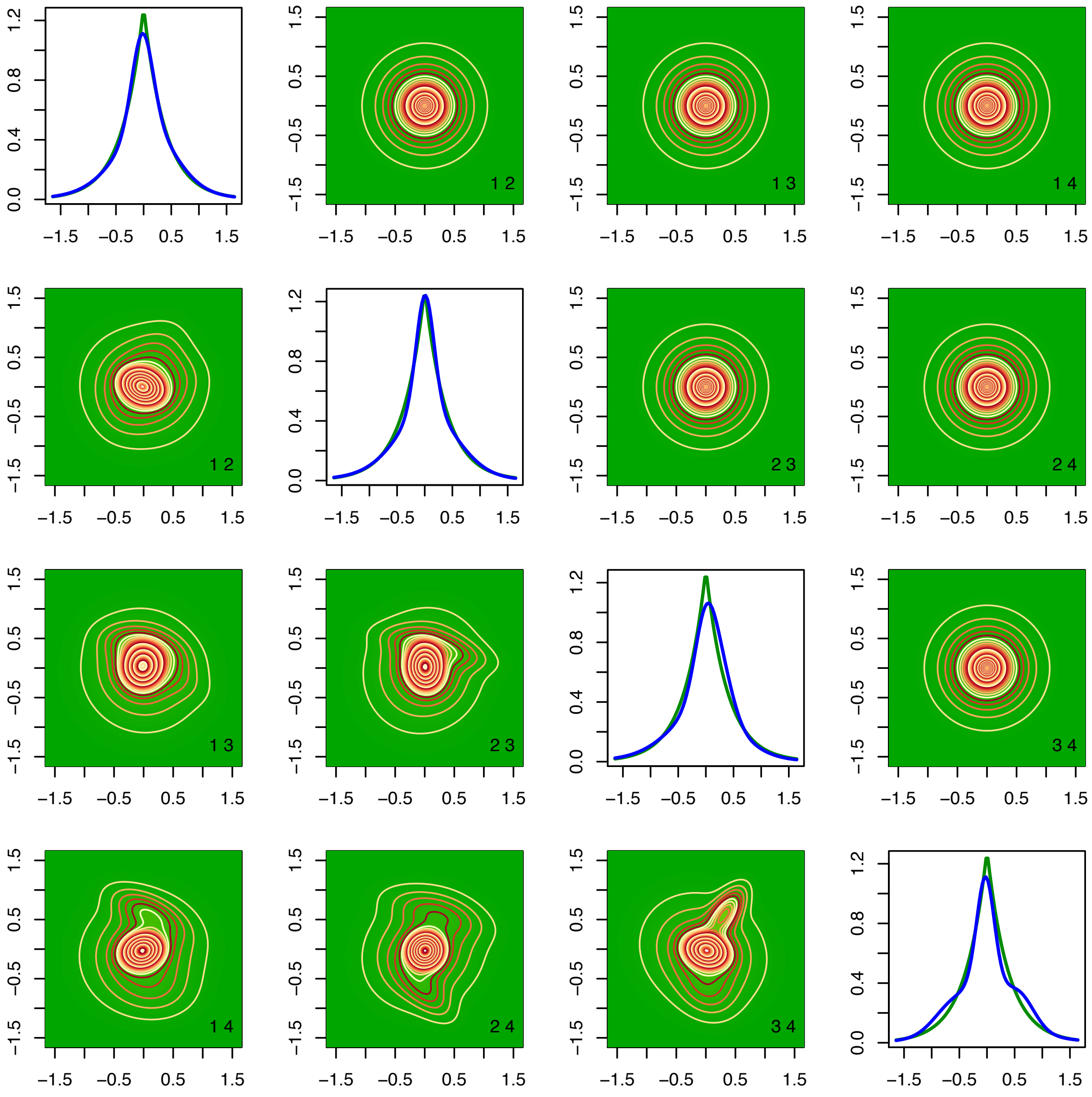


Figure S.19: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MIW (mixtures with inverse Wishart priors) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 and Section S.9 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities $f_{\epsilon_i,\epsilon_j}$ are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

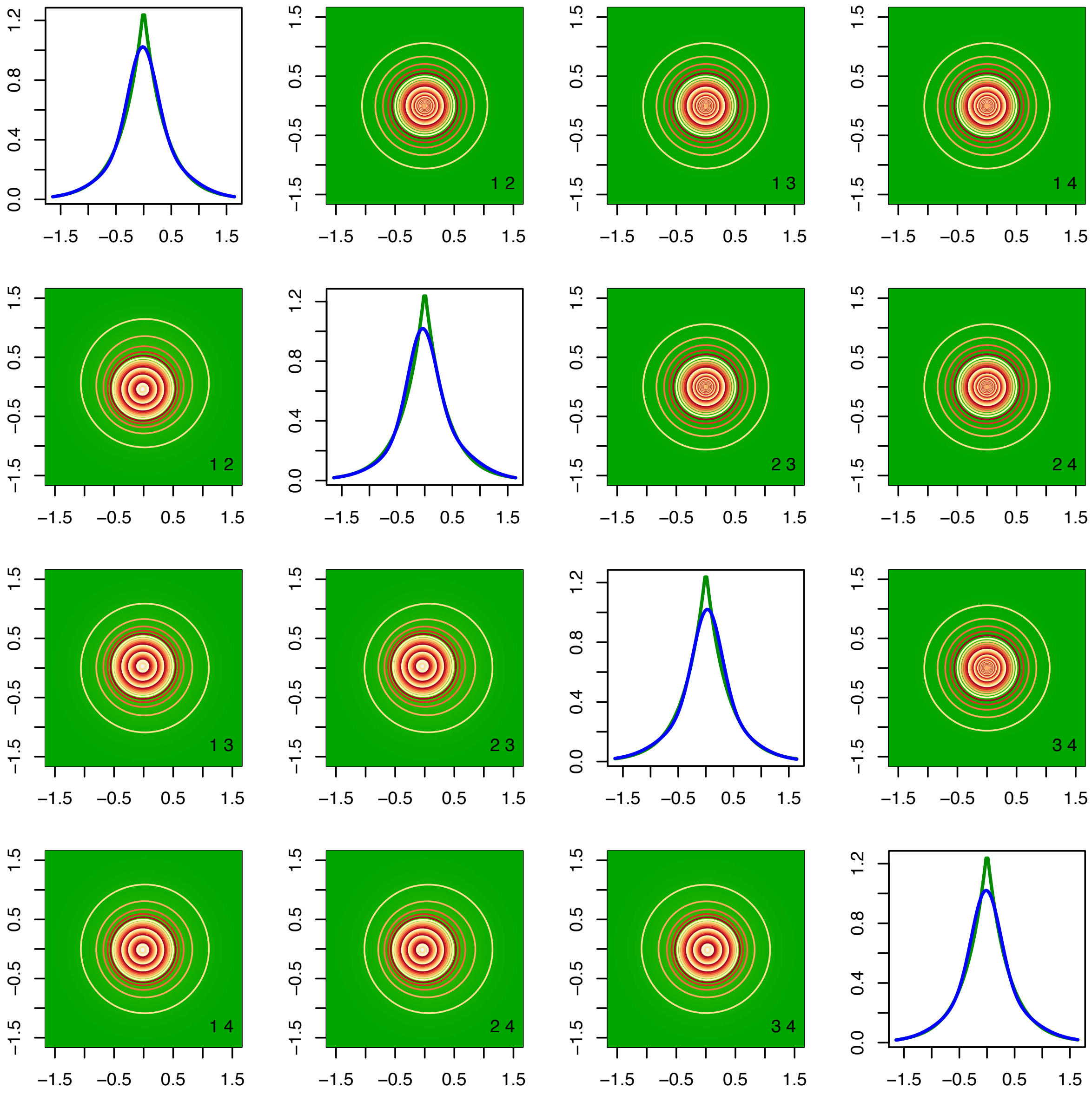


Figure S.20: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MLFA$_2$ (mixtures of latent factor analyzers) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 6 and Section S.9 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The numbers i, j at the bottom right corners of the off-diagonal panels show that the marginal densities $f_{\epsilon_i,\epsilon_j}$ are plotted in those panels. The diagonal panels show the true (lighter shaded green lines) and the estimated (darker shaded blue lines) one dimensional marginals. The figure is in color in the electronic version of this article.

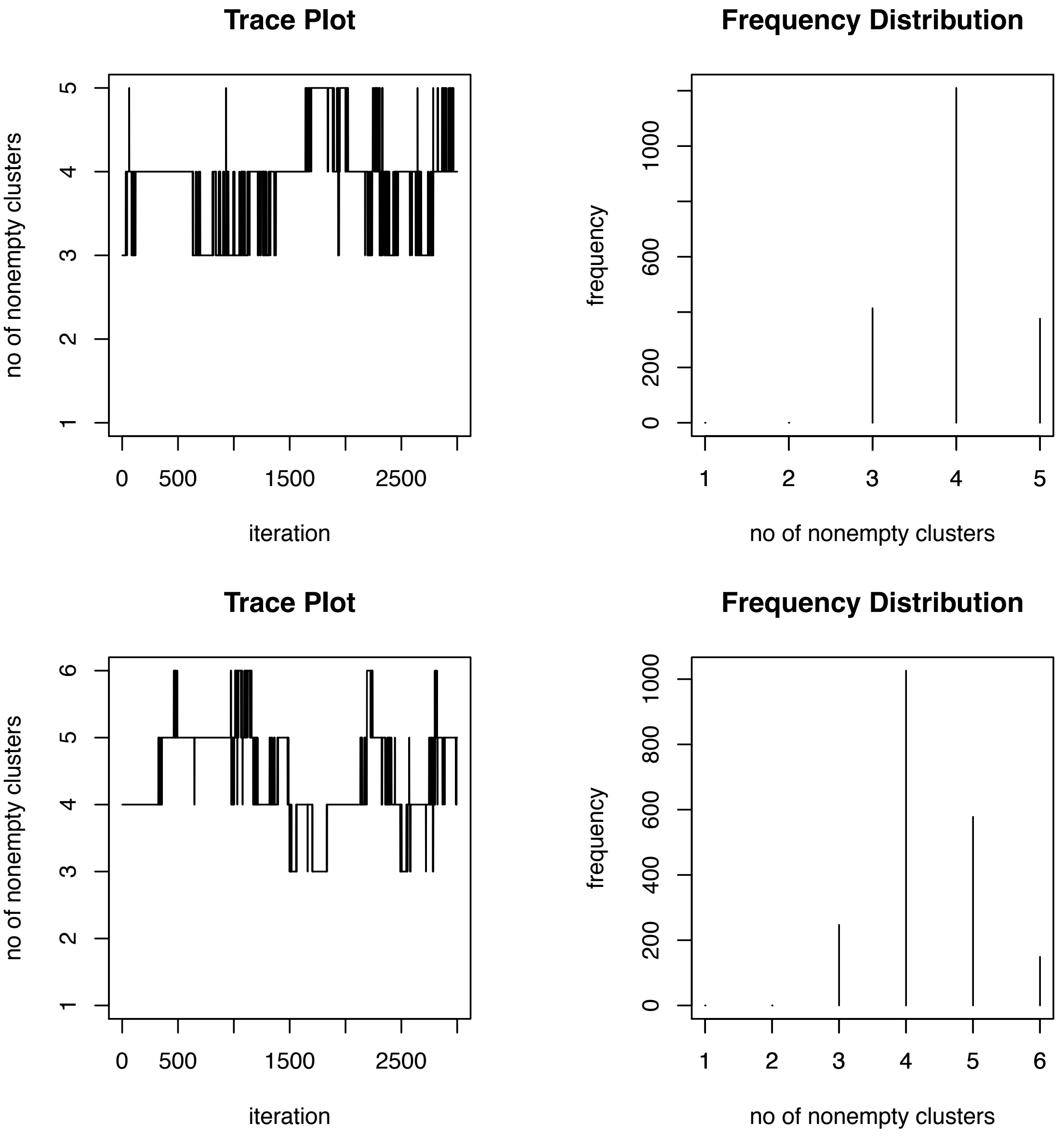


Figure S.21: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 6 and Section S.9 for additional details. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 5$ and $K_{\boldsymbol{\epsilon}} = 6$, respectively.

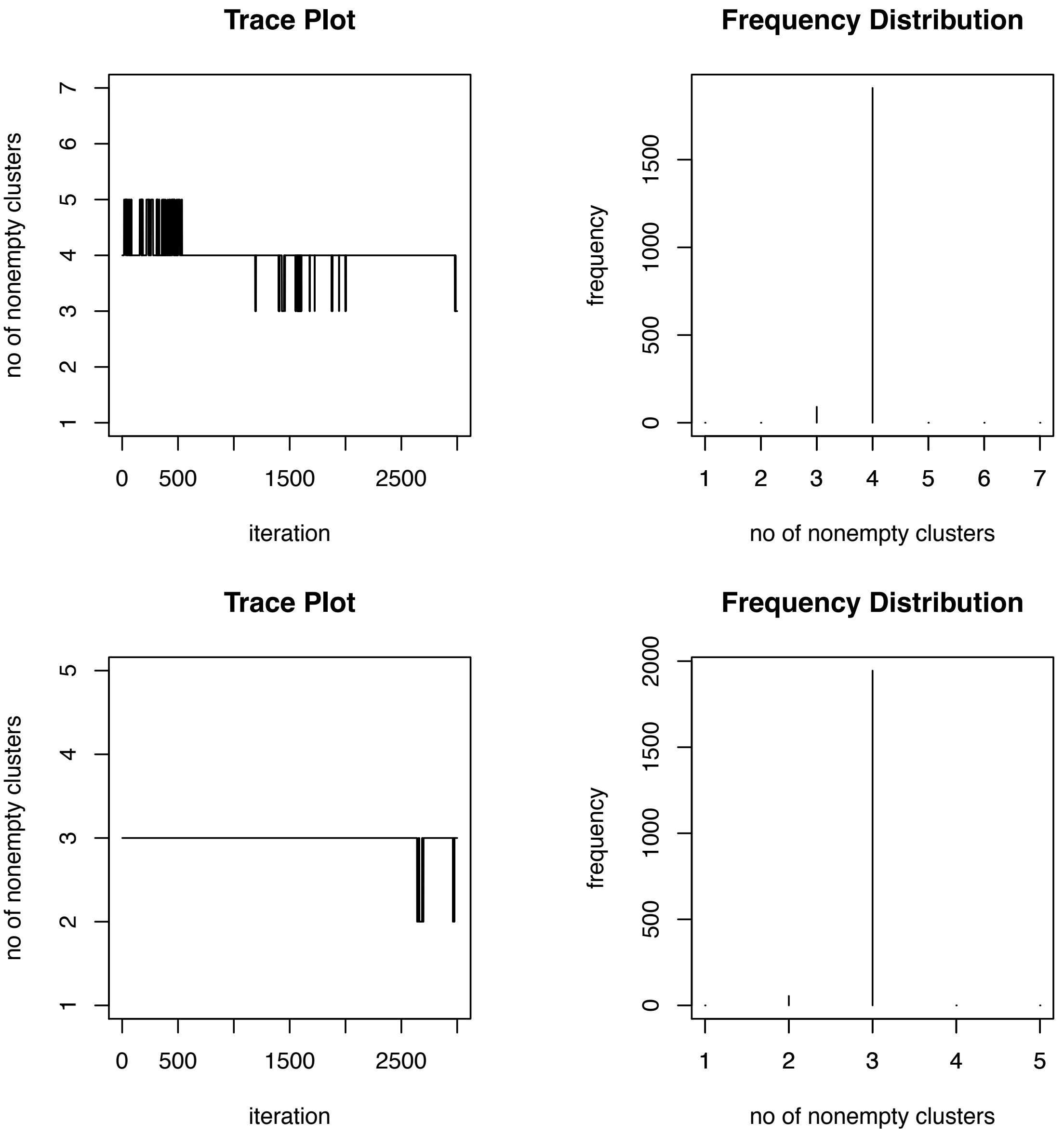
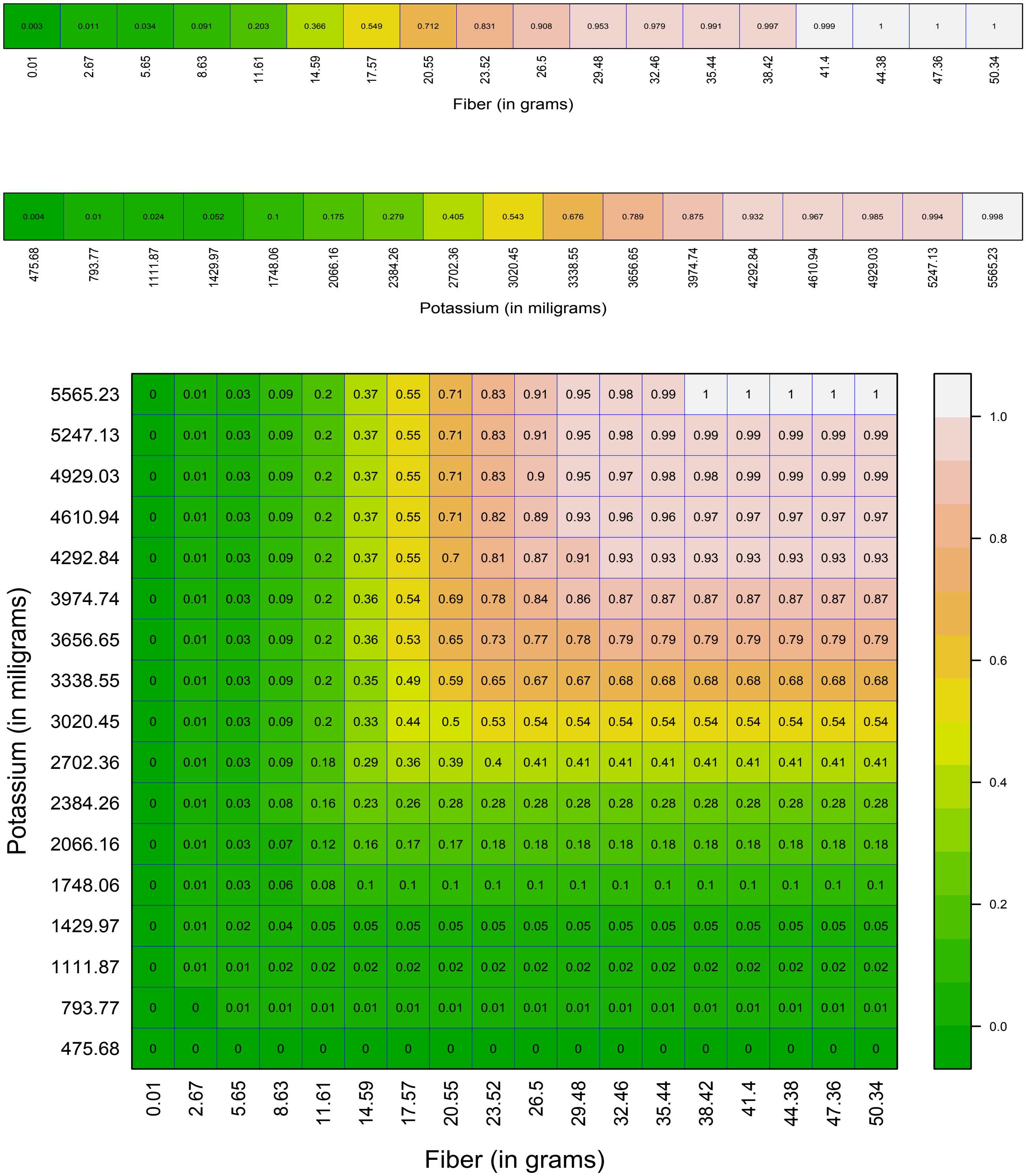


Figure S.22: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA$_2$ (mixtures of latent factor analyzers) method for conditionally heteroscedastic multivariate Laplace ($f_{\boldsymbol{\epsilon}}^{(4)}$) distributed measurement errors with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 6 and Section S.9 for additional details. The upper panels are for the $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$ were kept fixed at $K_{\mathbf{X}} = 7$ and $K_{\boldsymbol{\epsilon}} = 5$, respectively.

S.10 Potential Impact on Nutritional Epidemiology

The joint distribution of long-term average intakes of different dietary components allows nutritionists to study the dietary habits of the population of interest in fine detail. The plots of pairwise marginal distributions presented in Figure 8, for instance, provide detailed information on the joint consumption patterns of different pairs of dietary components. While such graphical summaries of the joint distributions may not be available for more than two components, numerical summaries of the joint distribution can provide answers to important questions such as what proportion of the population consume certain dietary components above, between or below certain amounts etc. The last question is particularly important as it relates to the proportion of the population that are deficient in certain dietary components. Focusing again on a two-dimensional case for illustration, namely Fiber and Potassium, Figure S.23 below shows their marginal and joint cumulative distribution function (CDF) on a set of grid points from which such proportions can be readily obtained. Dietary components are often reported in different measurement units. The figures presented in Section 7 are based on a linear scale transformation $W_{ij\ell} = 20\times\{W_{ij\ell,obs}-W_{ij\ell,obs,min}\}/\{W_{ij\ell,obs,max}-W_{ij\ell,obs,min}\}$ so that the $W_{ij\ell}$ for different components are unitless and fall between 0 and 20 units. Figure S.23 report the marginal and the joint CDF of fiber and potassium on a set of grid points in their original measurement units. We can readily see that, considered jointly, approximately 59% of adult Americans consume less than 20.55 grams of fiber and 3338.55 milligrams of potassium, whereas the corresponding marginal values are 71.2% and 67.6%, respectively.

The focus of the nutritional epidemiology examples considered in this article were on the estimation of joint consumption patterns of a set of regularly consumed dietary components whose reported intakes were all continuously measured. In contrast, for dietary components that are consumed episodically, the reported intakes equal zero on non-consumption days, and are positive on consumption days. The methodology developed in this article paves the way to more sophisticated deconvolution methods that can accommodate such zero inflated data. We are pursuing this problem as the subject of a separate study, with promising preliminary results. This will be a crucial step forward towards providing a highly flexible statistical framework for estimating the distribution of the U.S. Department of Agriculture's Healthy Eating Index (HEI, www.cnpp.usda.gov/HealthyEatingIndex.htm). HEI is a measure of diet quality that involves six episodically and seven regularly consumed dietary components and is used to assess compliance with the U.S. Dietary Guidelines for Americans (www.health.gov/dietaryguidelines) and monitor changes in dietary patterns. Efficient estimation of the distribution of HEI will allow nutritionists to answer public health questions that have important policy implications. We expect successful implementation of our methods to eventually replace the currently popular NCI method (www.riskfactor.cancer.gov/diet/usualintakes/method.html) for estimation of HEI.



Fiber (in grams)	0.01	2.67	5.65	8.63	11.61	14.59	17.57	20.55	23.52	26.5	29.48	32.46	35.44	38.42	41.4	44.38	47.36	50.34
	0.003	0.011	0.034	0.091	0.203	0.366	0.549	0.712	0.831	0.908	0.953	0.979	0.991	0.997	0.999	1	1	1

Potassium (in miligrams)	475.68	793.77	1111.87	1429.97	1748.06	2066.16	2384.26	2702.36	3020.45	3338.55	3656.65	3974.74	4292.84	4610.94	4929.03	5247.13	5565.23
	0.004	0.01	0.024	0.052	0.1	0.175	0.279	0.405	0.543	0.676	0.789	0.875	0.932	0.967	0.985	0.994	0.998

Potassium (in miligrams) \ Fiber (in grams)	0.01	2.67	5.65	8.63	11.61	14.59	17.57	20.55	23.52	26.5	29.48	32.46	35.44	38.42	41.4	44.38	47.36	50.34
5565.23	0	0.01	0.03	0.09	0.2	0.37	0.55	0.71	0.83	0.91	0.95	0.98	0.99	1	1	1	1	1
5247.13	0	0.01	0.03	0.09	0.2	0.37	0.55	0.71	0.83	0.91	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.99
4929.03	0	0.01	0.03	0.09	0.2	0.37	0.55	0.71	0.83	0.9	0.95	0.97	0.98	0.98	0.99	0.99	0.99	0.99
4610.94	0	0.01	0.03	0.09	0.2	0.37	0.55	0.71	0.82	0.89	0.93	0.96	0.96	0.97	0.97	0.97	0.97	0.97
4292.84	0	0.01	0.03	0.09	0.2	0.37	0.55	0.7	0.81	0.87	0.91	0.93	0.93	0.93	0.93	0.93	0.93	0.93
3974.74	0	0.01	0.03	0.09	0.2	0.36	0.54	0.69	0.78	0.84	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87
3656.65	0	0.01	0.03	0.09	0.2	0.36	0.53	0.65	0.73	0.77	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79
3338.55	0	0.01	0.03	0.09	0.2	0.35	0.49	0.59	0.65	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.68	0.68
3020.45	0	0.01	0.03	0.09	0.2	0.33	0.44	0.5	0.53	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
2702.36	0	0.01	0.03	0.09	0.18	0.29	0.36	0.39	0.4	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41
2384.26	0	0.01	0.03	0.08	0.16	0.23	0.26	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
2066.16	0	0.01	0.03	0.07	0.12	0.16	0.17	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18
1748.06	0	0.01	0.03	0.06	0.08	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
1429.97	0	0.01	0.02	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1111.87	0	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
793.77	0	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
475.68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure S.23: Results for Fiber and Potassium in their commonly used measurement units. The top two panels show their marginal cumulative distribution functions. The bottom panel shows their joint cumulative distribution function for a set of grid points. The figure is in color in the electronic version of this article.

Additional References

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Annals of Statistics*, 40, 206-237.

Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *Annals of Statistics*, 39, 2557-2584.

Carroll, R. J., Chen X. and Hu, Y. (2010). Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of Nonparametric Statistics*, 22, 379-399.

Castillo, I. and Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Annals of Statistics*, 42, 1941-1969.

de Boor, C. (2000). *A Practical Guide to Splines*. New York: Springer.

d'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27, 460-471.

Eltoft, T., Kim, T. and Lee, T. W. (2006). On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13, 300-303.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics*, 20, 1412-1425.

Ferguson, T. F. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.

Fraley, C. and Raftery, A. E. (2007). Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software*, 18, 1-13.

Ghosh, J. K. and Ramamoorthi, R. V. (2010). *Bayesian Nonparametrics*. New York: Springer.

Goldberg, R. R . (1961). *Fourier transforms*. Volume 32. London: Cambridge.

Green, J. P., Latuszynski, K. Pereyra, M. and Roberts, C. P. (2015). Bayesian computation: summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25, 835-862.

Hastie, D. I., Liverani, S. and Richrdson, S. (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25, 1023-1037.

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11, 508-532.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 371-390.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, 30, 269-283.

Johnstone, I. M. (2010). High dimensional Bernstein-von Mises: simple examples. *Institute of Mathematical Statistics Collections*, 6, 87-98.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge: Cambridge University Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249-265.

Norets, A. and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168, 332-346.

Pati, D. and Dunson, D. (2013). Bayesian nonparametric regression with varying residual density. *Annals of the Institute of Statistical Mathematics*, 66, 1-13.

Pelenis, J. (2014). Bayesian Regression with Heteroscedastic Error Density and Parametric Mean Function. *Journal of Econometrics*, 178, 624-638.

Rocke, D. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8, 557-569.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behavior of the posterior distribution in overfitted mixture models *Journal of the Royal Statistical Society, Series B*, 73, 689-710.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Spokoiny, V. (2013). Bernstein-von Mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430.*