# Scalable multiscale density estimation

Supplementary Material

## Contents



## 1 Proof of Theorem 1

The log-posterior of GEODE, up to a additive constant, is as follows

$$\mathcal{L} = -\frac{N}{2}\big\{\ln|\boldsymbol{C}| + \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S}) - \tilde{a}\ln(\sigma^2) + \tilde{b}\sigma^{-2}\big\},$$

where $\tilde{a} = \frac{2a_\sigma+2}{N}$ and $\tilde{b} = \frac{2b_\sigma}{N}$. Consider the empirical Bayes problem:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{W}}) = \arg\max_{\boldsymbol{\mu},\boldsymbol{W}} \big[\max_{\sigma^2,\boldsymbol{\Sigma}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2, \boldsymbol{\Sigma})\big]. \tag{1}$$

**Theorem 1.** *Let $\lambda_1, \ldots, \lambda_D$ be the eigenvalues of $\boldsymbol{S}$ ordered descendingly, define $e_k = (D-d)\lambda_k - \sum_{j=k}^{k+D-d-1}\lambda_j$ for all $k \leqslant d+1$ and define $q = \sum_{j=d+1}^{D}\lambda_j - (D-d)\lambda_D$. Suppose*

*Condition 1:* $d < rank(\boldsymbol{S})$

*Condition 2:* $(a_\sigma + 1)\lambda_D \leqslant \frac{N}{2} q$

*Condition 3: For all* $e_k > 0$, $b_\sigma < \frac{N}{2} e_k$.

*Then*

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}},\ \hat{\boldsymbol{W}} = \boldsymbol{U}_d$$

*solves (1), where the* $d$ *column vectors in the* $D \times d$ *matrix* $\boldsymbol{U}_d$ *are the* $d$ *leading eigenvectors of* $\boldsymbol{S}$.

*Proof.* The proof can be split into two parts.

**Part 1:** $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ is the stationary point of $\mathcal{L}$.

From standard matrix differentiation results:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} &= -N\Big\{ \sum_{i=1}^{N} \boldsymbol{C}^{-1}(\boldsymbol{\mu} - \boldsymbol{y}_i) \Big\} \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} &= -\frac{N}{2}\Big\{ 2\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} - 2\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} \Big\}
\end{aligned}$$

Solving for $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = 0$ gives $\boldsymbol{\mu} = \bar{\boldsymbol{y}}$. Solving for $\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = 0$ gives

$$\begin{aligned}
&\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{\Sigma} \\
\Leftrightarrow &\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} = \boldsymbol{W} \qquad (2)
\end{aligned}$$

Neither of the two trivial solutions to (2), $\boldsymbol{W} = \boldsymbol{0}$ and $\boldsymbol{C} = \boldsymbol{S}$ maximizes $\mathcal{L}$ and hence will not be discussed. The left solution corresponds to a $\boldsymbol{W}$ such that $\boldsymbol{W} \neq \boldsymbol{0}$ and $\boldsymbol{C} \neq \boldsymbol{S}$. With the fact that column vectors of $\boldsymbol{W}$ are orthonormal, and using the result from Henderson and Searle (1981), we have

$$\begin{aligned}
&\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} = \boldsymbol{W} \\
\Leftrightarrow &\boldsymbol{S}\big[\sigma^{-2}\boldsymbol{I} - \sigma^{-2}\boldsymbol{W}(\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{W}^\top\big]\boldsymbol{W} = \boldsymbol{W} \\
\Leftrightarrow &\boldsymbol{S}\big[(\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma}) - \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top\big]\boldsymbol{W} = \sigma^2(\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma})\boldsymbol{W} \\
\Leftrightarrow &\boldsymbol{S}\boldsymbol{W} = (\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma})\boldsymbol{W} \qquad (3)
\end{aligned}$$

Equation (3) implies that each column of $\boldsymbol{W}$ must be an eigenvector of $\boldsymbol{S}$, with corresponding eigenvalues $\gamma_j = \sigma^2 + \alpha_j^2$. Note that this also implies that $\sigma^2 \leqslant \gamma_j$ for $j = 1, \ldots, d$.

Now we check if $\frac{\partial \mathcal{L}}{\partial \alpha_j^2}|_{\alpha_j^2 = \gamma_j - \sigma^2} = 0$, which will complete the proof that $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ is the stationary point of $\mathcal{L}$.

We substitute stationary point of $\boldsymbol{W}$ into $\mathcal{L}$ to give

$$\begin{aligned}\mathcal{L} = -\frac{N}{2}\Big\{(D-d)\ln(\sigma^2) + \sum_{j=1}^{d}\ln(\sigma^2+\alpha_j^2) + \frac{1}{\sigma^2}\sum_{j=1}^{D}\gamma_j \\ -\frac{1}{\sigma^2}\sum_{j=1}^{d}\frac{\gamma_j\alpha_j^2}{\sigma^2+\alpha_j^2} + \tilde{a}\ln(\sigma^2) + \tilde{b}\sigma^{-2}\Big\}.\end{aligned} \tag{4}$$

From (4) one can easily check $\frac{\partial\mathcal{L}}{\partial\alpha_j^2}|_{\alpha_j^2=\gamma_j-\sigma^2} = 0$, for $j = 1,\ldots,d$.

**Part 2:** Show $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{W}} = \boldsymbol{U}_p$ maximizes $\mathcal{L}$.

Matrix $\boldsymbol{W}$ may contain any of the eigenvectors of $\boldsymbol{S}$. To figure out when is $\mathcal{L}$ maximized, we substitute stationary point of $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$ into $\mathcal{L}$ to give

$$\begin{aligned}\mathcal{L} = -\frac{N}{2}\Big\{\sum_{j=1}^{d}\ln(\gamma_j) + \frac{1}{\sigma^2}\sum_{j=d+1}^{D}\gamma_j + (D-d)\ln\sigma^2 + d \\ + \tilde{a}\ln\sigma^2 + \tilde{b}\frac{1}{\sigma^2}\Big\},\end{aligned} \tag{5}$$

where $\gamma_1,\ldots,\gamma_d$ are the eigenvalues corresponding to the eigenvectors 'retained' in $\boldsymbol{W}$ and $\gamma_{d+1},\ldots,\gamma_D$ are those 'discarded'. Here we slightly abuse notations: we use $\lambda_1,\ldots,\lambda_D$ as the eigenvalues of $\boldsymbol{S}$ ordered descendingly. We use $\gamma_1,\ldots,\gamma_D$ also as the eigenvalues of $\boldsymbol{S}$ but with the first $d$ corresponding to the stationary point $\boldsymbol{W}$. Note that $\gamma_j$'s are not necessarily ordered.

Maximizing (5) w.r.t. $\sigma^2$ gives

$$\sigma^2 = \frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d+1}^{D}\gamma_j + \tilde{b}\Big) > 0. \tag{6}$$

When $\lambda_D > 0$, with condition 2, it is easy to check that

$$\frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d}^{D-1}\lambda_j + \tilde{b}\Big) > \lambda_D$$

Since $\sigma^2 \leqslant \gamma_j$, for $j = 1,\ldots,d$, we know immediately that $\lambda_D$ has to be discarded. If $\lambda_D = 0$, it is obvious that it also has to be discarded since $\sigma^2 > 0$. Note that Condition 1 ensures the existence of $b_\sigma$ that satisfies condition 3. Condition 3 ensures the existence of the stationary point to $\mathcal{L}$ since

$$\frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d+1}^{D}\lambda_j + \tilde{b}\Big) < \lambda_d,$$

which means that we at least have one stationary point solution.

Substituting $\sigma^2$ w.r.t. (6) gives

$$\begin{aligned}\mathcal{L} = -\frac{N}{2}\Big\{ \sum_{j=1}^{d} \ln(\gamma_j) + D + \tilde{a} + \\ (D-d+\tilde{a}) \ln\Big[\frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d+1}^{D} \gamma_j + \tilde{b}\Big)\Big]\Big\}.\end{aligned} \tag{7}$$

When all eigenvalues are non-zero, with the fact that $\sum_{j=1}^{D} \ln(\gamma_j)$ is a constant, maximizing (7) is equivalent to minimizing the following quantity

$$E = \ln\Big[\frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d+1}^{D} \gamma_j + \tilde{b}\Big)\Big] - \frac{1}{D-d+\tilde{a}}\Big(\sum_{j=d+1}^{D} \ln(\gamma_j) + \tilde{b}\Big). \tag{8}$$

When there are zero eigenvalues, we simply ignore these zero eigenvalues and consider only the non-zeros ones and every statement in this proof holds.

It turns out that all is required for (8) to be minimized is that $\gamma_{d+1}, \ldots, \gamma_D$ *are adjacent within the spectrum of the ordered eigenvalues of* $\boldsymbol{S}$. To see this, the first order derivative of $E$ is given by

$$\frac{\partial E}{\partial \gamma_j} = \frac{1}{\sum_{j=d+1}^{D} \gamma_j + \tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_j}. \tag{9}$$

Suppose $\gamma_{d+1}, \ldots, \gamma_D$ minimizes (8). Without loss of generality, we assume $\gamma_{d+1} \geqslant \ldots \geqslant \gamma_D$. From the previous discussion we know that $\gamma_D = \lambda_D$. Define $c = \sum_{j=d+2}^{D-1} \gamma_j$ when $D - d \geqslant 2$ We don't need to discuss the case $D - d = 1$. Since if that is the case then $\gamma_1, \ldots, \gamma_d$ must be the $d$ leading eigenvalues because $\lambda_D$ has to be discarded.

From (9) we immediately have

$$\begin{aligned}\frac{\partial E}{\partial \gamma_{d+1}} &= \frac{1}{\gamma_{d+1} + \gamma_D + c + \tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_{d+1}}. \\ \frac{\partial E}{\partial \gamma_D} &= \frac{1}{\gamma_D + \gamma_{d+1} + c + \tilde{b}} - \frac{1}{(D-d+\tilde{a})\gamma_D}.\end{aligned}$$

From Condition 3 it is easy to check that

$$\gamma_D + \gamma_{d+1} + c + \tilde{b} < (D-d)\lambda_{d+1}.$$

hence $\frac{\partial E}{\partial \gamma_{d+1}} > 0$. Similarly, with condition 2 one can also check that $\frac{\partial E}{\partial \gamma_D} < 0$. It can also be checked that for any $\lambda_{d+1}$ such that $\lambda_{d+1} > \frac{\lambda_D + c + \tilde{b}}{D-d+\tilde{a}-1}$, $\frac{\partial E}{\partial \gamma_{d+1}} > 0$ holds. And for any $\lambda_D$ such that $\lambda_D < \frac{\lambda_{d+1} + c + \tilde{b}}{D-d+\tilde{a}-1}$, $\frac{\partial E}{\partial \gamma_D} < 0$ holds. Moreover, $\frac{\lambda_{d+1} + c + \tilde{b}}{D-d+\tilde{a}-1} > \frac{\lambda_D + c + \tilde{b}}{D-d+\tilde{a}-1}$. Hence $\gamma_{d+1}, \ldots, \gamma_D$ have to be adjacent within the spectrum of the ordered eigenvalues of $\boldsymbol{S}$. Otherwise, if there is a $\gamma$ in between such that $\gamma_D < \gamma < \gamma_{d+1}$ then either $\frac{\partial E}{\partial \gamma_D}|_{\gamma_D = \gamma} < 0$ or $\frac{\partial E}{\partial \gamma_{d+1}}|_{\gamma_{d+1} = \gamma} > 0$ must be

true. Hence either replacing $\gamma_D$ or $\gamma_{d+1}$ with $\gamma$ will further decrease $\mathcal{L}$, which contradicts the assumption that $\gamma_{d+1}, \ldots, \gamma_D$ minimizes $\mathcal{L}$.

Coupled with the fact that $\gamma_D = \lambda_D$, we have shown that the $D-d$ smallest eigenvalues minimizes (8) hence $\mathcal{L}$ is maximized if $\gamma_1, \ldots, \gamma_d$ are the $d$ leading eigenvalues of $\boldsymbol{S}$. Hence we have

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$$

and substituting $\boldsymbol{\mu} = \bar{\boldsymbol{y}}$ into $\boldsymbol{S}$ we have $\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{Y}^\top$ hence we have

$$\hat{\boldsymbol{W}} = \boldsymbol{U}_d$$

□

# 2 Likelihood of GEODE

Let introduce sufficient statistics $A_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})^\top (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})$ and $\boldsymbol{Z}_i = \hat{\boldsymbol{W}}^\top (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})$, with $Z_i^{(j)}$ denoting the $j$th element of $\boldsymbol{Z}_i$. We then apply a random variable transformation $u_j = (1 + \sigma^{-2}\alpha_j^2)^{-1}$, for $j = 1, \ldots, d$. The likelihood of GEODE is then

$$f(\boldsymbol{y}_i) \propto (\sigma^2)^{-D/2} \prod_{j=1}^{d} u_j^{1/2} \exp\left\{ -\frac{1}{2}\sigma^{-2} \times \left[ A_i - \sum_{j=1}^{d} (1-u_j)(Z_i^{(j)})^2 \right] \right\}. \quad (10)$$

which can be derived using the following two facts.

**Fact 1.** *$\boldsymbol{\Sigma} = diag(\alpha_1^2, \ldots, \alpha_d^2)$ is a $d \times d$ matrix with all diagonal entries larger than 0, $\boldsymbol{W}$ is a $D \times d$ matrix with orthonormal column vectors, we have,*

$$(\sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^{-1} = \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{W}\tilde{\boldsymbol{\Sigma}}\boldsymbol{W}^\top,$$

*where* $\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\frac{\alpha_1^2}{1+\sigma^{-2}\alpha_1^2}, \frac{\alpha_2^2}{1+\sigma^{-2}\alpha_2^2}, \ldots, \frac{\alpha_d^2}{1+\sigma^{-2}\alpha_d^2})$.

*Proof.* By the orthonormality of the $\boldsymbol{W}$, we have $\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}$. And by the matrix inversion formula (Henderson and Searle, 1981),

$$\begin{aligned}
(\sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^{-1} &= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{W}(\boldsymbol{I} + \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{W}(\boldsymbol{I} + \sigma^{-2}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{W}^\top \\
&= \sigma^{-2}\boldsymbol{I} - \sigma^{-4}\boldsymbol{W}\tilde{\boldsymbol{\Sigma}}\boldsymbol{W}^\top
\end{aligned}$$

□

**Fact 2.** *Under the same setting of Fact 1, we have*

$$|\sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top|^{-1/2} = (\sigma^2)^{-D/2} \prod_{j=1}^{d} \left(\frac{1}{1+\sigma^{-2}\alpha_j^2}\right)^{1/2}.$$

*Proof.* By Schur's formula,

$$\begin{aligned}|\sigma^2\boldsymbol{I}+\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top|^{-1/2} &= (\sigma^2)^{-D/2}|\boldsymbol{I}_D+\sigma^{-2}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top|^{-1/2}\\ &= (\sigma^2)^{-D/2}|\boldsymbol{I}_d+\sigma^{-2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{\Sigma}^{1/2}|^{-1/2}\\ &= (\sigma^2)^{-D/2}|\boldsymbol{I}_d+\sigma^{-2}\boldsymbol{\Sigma}|\\ &= (\sigma^2)^{-D/2}\prod_{j=1}^{d}(\frac{1}{1+\sigma^{-2}\alpha_j^2})^{1/2}\end{aligned}$$

□

# 3 Posterior Conditional Derivation

Based on the likelihood (10) of GEODE, the derivation details of the conditional posterior distributions are given as follows:
For $\sigma^2$,

$$\begin{aligned}&p(\sigma^{-2}|-)\\ &\propto(\sigma^{-2})^{a_\sigma-1}\exp(-b_\sigma\sigma^{-2})\prod_{i=1}^{N}(\sigma^2)^{-D/2}\exp\Big\{-\frac{1}{2}\sigma^{-2}(A_i-\sum_{j=1}^{d}(1-u_j)(Z_i^{(j)})^2)\Big\}\\ &\propto(\sigma^{-2})^{DN/2+a_\sigma-1}\exp\Big\{-\sigma^{-2}[\frac{1}{2}\sum_{i=1}^{N}(A_i-\sum_{j=1}^{d}(1-u_j)(Z_i^{(j)})^2)+b_\sigma]\Big\}.\end{aligned}$$

For $u_j$, for $j=1,\ldots,d$,

$$\begin{aligned}&p(u_j|-)\\ &\propto\prod_{i=1}^{N}u_j^{1/2}\exp\Big\{-\frac{1}{2}\sigma^{-2}u_j(Z_i^{(j)})^2\Big\}u_j^{\prod_{k=1}^{j}\tau_k-1}\exp\{-u_j\}\mathbb{1}_{(0,1)}\\ &\propto u_j^{\prod_{k=1}^{j}\tau_k+N/2-1}\exp\Big\{-[1+\frac{1}{2}\sigma^{-2}\sum_{i=1}^{N}(Z_i^{(j)})^2]u_j\Big\}\mathbb{1}_{(0,1)}.\end{aligned}$$

For $\tau_j$, for $j=1,\ldots,d$,

$$\begin{aligned}&p(\tau_j|-)\\ &\propto(\prod_{k>j-1}u_k)^{\tau_k}\exp\{-a_\tau\tau_j\}\mathbb{1}_{[1,\infty)}\\ &\propto\exp\Big\{-[a_\tau-ln(\prod_{k>j-1}u_k)]\tau_j\Big\}\mathbb{1}_{[1,\infty)}\end{aligned}$$

# 4 Proof of Proposition 1 and Corollary 1

Notations $\boldsymbol{y}_M$ and $\boldsymbol{y}_O$ are introduced as the missing part and the observed part of $\boldsymbol{y}$ respectively. Let $\boldsymbol{\mu}_M$ and $\boldsymbol{W}_M$ denote the parts of $\boldsymbol{\mu}$ and $\boldsymbol{W}$ corresponding

to $\boldsymbol{y}_M$, and let $\boldsymbol{\mu}_O$ and $\boldsymbol{W}_O$ denote the parts corresponding to $\boldsymbol{y}_O$. The following proposition enables efficient sampling from the conditional posterior distribution $p(\boldsymbol{y}_M|\boldsymbol{y}_O, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ denotes all the unknown parameters in the model.

**Proposition 1.** *Introduce augmented data $\boldsymbol{\eta} \in \Re^d$ such that $(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\Theta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\eta}, \sigma^2\boldsymbol{I})$ and $(\boldsymbol{\eta}|\boldsymbol{\Theta} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Then we have the conditional distribution with $\boldsymbol{\eta}$ marginalized out equal $(\boldsymbol{y}|\boldsymbol{\Theta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top, \sigma^2\boldsymbol{I})$. Furthermore, we have*

$$\begin{aligned}\boldsymbol{\eta}|\boldsymbol{y}_O, \boldsymbol{\Theta} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_\eta, \hat{\boldsymbol{C}}_\eta),\\ \boldsymbol{y}_M|\boldsymbol{\eta}, \boldsymbol{y}_O, \boldsymbol{\Theta} &\sim \mathcal{N}(\boldsymbol{\mu}_M + \boldsymbol{W}_M\boldsymbol{\eta}_i, \sigma^2\boldsymbol{I}),\end{aligned}$$

*where $\hat{\boldsymbol{C}}_\eta = \left(\boldsymbol{\Sigma}\boldsymbol{W}_O^\top\boldsymbol{W}_O/\sigma^2 + \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\mu}}_\eta = \hat{\boldsymbol{C}}_\eta\boldsymbol{W}_O^\top(\boldsymbol{y}_O - \boldsymbol{\mu}_O)/\sigma^2$.*

*Proof.* The proposition can be easily proved using Bayes rule. The joint density of $(\boldsymbol{y}_O, \boldsymbol{y}_M, \boldsymbol{\eta}|\boldsymbol{\Theta})$ is given by

$$\begin{aligned}&p(\boldsymbol{y}_O, \boldsymbol{y}_M, \boldsymbol{\eta}|\boldsymbol{\Theta})\\ \propto &\exp\left\{-\frac{\|\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\mu}\|_2}{2\sigma^2} - \frac{\boldsymbol{\eta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}}{2}\right\}\\ \propto &\exp\left\{-\frac{\|\boldsymbol{y}_M - \boldsymbol{W}_M\boldsymbol{\eta} - \boldsymbol{\mu}_M\|_2}{2\sigma^2} - \frac{\|\boldsymbol{y}_O - \boldsymbol{W}_O\boldsymbol{\eta} - \boldsymbol{\mu}_O\|_2}{2\sigma^2} - \frac{\boldsymbol{\eta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}}{2}\right\}.\end{aligned}$$

Hence the conditional density $(\boldsymbol{y}_M|\boldsymbol{\eta}, \boldsymbol{y}_O, \boldsymbol{\Theta})$ is given by

$$p(\boldsymbol{y}_M|\boldsymbol{\eta}, \boldsymbol{y}_O, \boldsymbol{\Theta}) \propto \exp\left\{-\frac{\|\boldsymbol{y}_M - \boldsymbol{W}_M\boldsymbol{\eta} - \boldsymbol{\mu}_M\|_2}{2\sigma^2}\right\}.$$

The marginal conditional density $(\boldsymbol{\eta}|\boldsymbol{y}_O, \boldsymbol{\Theta})$ is given by

$$\begin{aligned}&p(\boldsymbol{\eta}|\boldsymbol{y}_i^O, \boldsymbol{\Theta})\\ \propto &\int p(\boldsymbol{y}_M, \boldsymbol{\eta}|\boldsymbol{y}_O)\mathrm{d}\boldsymbol{y}_M\\ \propto &\exp\left\{\frac{\|\boldsymbol{y}_O - \boldsymbol{W}_O\boldsymbol{\eta}_i - \boldsymbol{\mu}_O\|_2}{2\sigma^2} - \frac{\boldsymbol{\eta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}}{2}\right\}.\end{aligned}$$

□

**Corollary 1.** *For any $\boldsymbol{\Theta}$, the following are true*

$$\begin{aligned}\boldsymbol{y}_O|\boldsymbol{\Theta}, \boldsymbol{\mu}_O, \boldsymbol{W}_O &\sim \mathcal{N}(\boldsymbol{\mu}_O, \boldsymbol{W}_O\boldsymbol{\Sigma}\boldsymbol{W}_O^\top + \sigma^2\boldsymbol{I}),\\ \boldsymbol{y}_M|\boldsymbol{y}_O, \boldsymbol{\Theta}, \boldsymbol{\mu}_O, \boldsymbol{W}_O &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{C}}_M),\end{aligned}$$

*where $\hat{\boldsymbol{\mu}}_M = \boldsymbol{\mu}_M + \boldsymbol{W}_M\hat{\boldsymbol{\mu}}_\eta$ and $\hat{\boldsymbol{C}}_M = \boldsymbol{W}_M\hat{\boldsymbol{C}}_\eta\boldsymbol{W}_M^\top + \sigma^2\boldsymbol{I}$.*

Corollary 1 is a direct result from Proposition 1 and the multivariate Gaussian theory.

Table 1: Dataset 1

| | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $\sigma^2$ |
|---|---|---|---|---|---|---|
| $\hat{R}$ | 0.9998 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 0.9998 |
| $n_{eff}$ | 675 | 658 | 626 | 671 | 637 | 651 |

# 5 Missing Data Imputation for mGEODE

Conditional on the membership $(s_i, h_i)$, the imputational strategies of nonlinear GEODE are exactly the same as those of the linear GEODE. With a slight abuse of notations, we denote the parts corresponding to $\boldsymbol{y}_O$ of $\boldsymbol{\mu}_{sh}$ and $\boldsymbol{W}_{sh}$ as $\boldsymbol{\mu}_{\boldsymbol{O}}$ and $\boldsymbol{W}_O$. Hence we only discuss the conditional posterior distribution of the membership variable given a partially observed $\boldsymbol{y}_O$, which is given as follows

$$\begin{aligned} &p(s_i = s, h_i = h|\boldsymbol{y}_O, \boldsymbol{\Theta}, \{\boldsymbol{\mu}_O, \boldsymbol{W}_O\}_L) \\ \propto &\pi_{s,h}\phi(\boldsymbol{y}_O; \boldsymbol{\mu}_O, \boldsymbol{W}_O\boldsymbol{\Sigma}_{sh}\boldsymbol{W}_{sh}^\top + \sigma_s^2\boldsymbol{I}) \end{aligned}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

# 6 METIS

To illustrate the dyadic clustering tree obtained from METIS, we simulated 3003 points first from a swissroll and then from a hemisphere. The results are visualized in Figure 1.

# 7 Convergence and Mixing of Gibbs Sampler

To diagnose the convergence of the proposed Gibbs sampler for GEODE, we follow the Gelman-Rubin diagnostic and ran two chains with over-dispersed random starts and calculate the potential scale reduction factor $\hat{R}$ for each random scalars. To diagnose the mixing, we pick one of the two chains and calculate the effective sample size $n_{eff}$ out of 2000 posterior samples for each random scalars. We use two simulated datasets to illustrate the superb convergence and mixing of the proposed Gibbs sampler. In the first dataset, we set $D = 10^4$, $p = 5$, $d = 10$ and $N = 500$. Both chains correctly select the first 5 dimensions. The diagnostic details are summarized in Table 1. In the second dataset, we set $D = 10^5$, $p = 10$, $d = 20$ and $N = 500$. Both chains again correctly select the first 10 dimensions. The diagnostic details are summarized in Table 2. As can be seen, all the $\hat{R}$'s are very close to 1, indicating good convergence. All the $n_{eff}$ is fairly large, indicating good mixing.

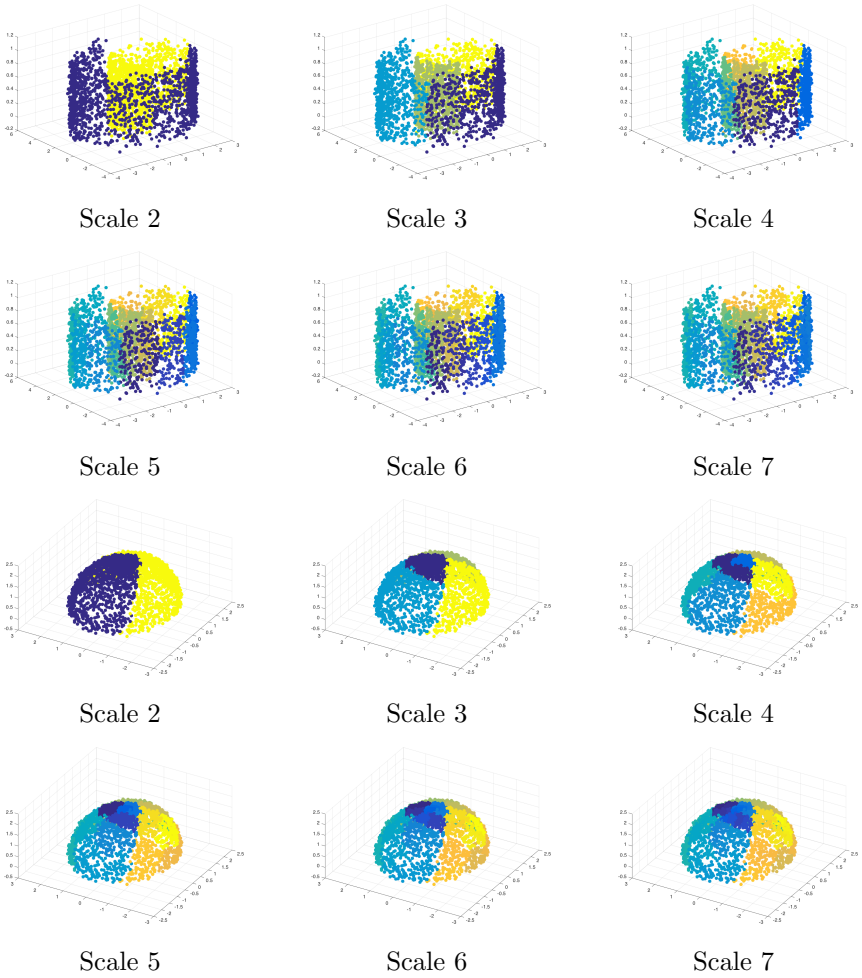


Figure 1: The dyadic clustering tree at each scale generated from METIS for synthetic data on a swissroll and on a hemisphere.

Table 2: Dataset 2

| | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $\sigma^2$ |
|---|---|---|---|---|---|---|
| $\hat{R}$ | 0.9998 | 0.9999 | 0.9998 | 0.9999 | 0.9999 | 0.9999 |
| $n_{eff}$ | 655 | 642 | 627 | 717 | 678 | 702 |
| | $u_6$ | $u_7$ | $u_8$ | $u_9$ | $u_{10}$ | |
| $\hat{R}$ | 1.0004 | 1.0006 | 1.0003 | 0.9999 | 1.0001 | |
| $n_{eff}$ | 686 | 728 | 719 | 660 | 674 | |

# References

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.