Supplementary Material for Hui et al., Semiparametric Regression using Variational Approximations

A Proofs and Derivations

A.1 Derivation of $\underline{\ell}_{\text{Norm}}(\boldsymbol{\Psi}, \boldsymbol{\xi})$

First note that $\int \boldsymbol{z}_i^\top \boldsymbol{\beta}\, h(\boldsymbol{\beta}|\boldsymbol{a}, \boldsymbol{A}) d\boldsymbol{\beta} = \boldsymbol{z}_i^\top \boldsymbol{a}$, which deals with the third term in $\ln\{f(y_i|\boldsymbol{\Psi}, \boldsymbol{\beta})\}$. Using the fact that $\boldsymbol{\beta}$ is multivariate normal with respect to the variational distribution, we obtain $\int \boldsymbol{\beta}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{\beta}\, h(\boldsymbol{\beta}|\boldsymbol{a}, \boldsymbol{A}) d\boldsymbol{\beta} = \boldsymbol{a}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{a} + \text{tr}\left(\boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{A}\right) = \boldsymbol{a}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{a} + \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i$. Combining the above results, we thus obtain $\int \ln\{f(y_i|\boldsymbol{\Psi}, \boldsymbol{\beta})\}\, h(\boldsymbol{\beta}|\boldsymbol{a}, \boldsymbol{A}) d\boldsymbol{\beta} = -(2\phi)^{-1} \left\{\tilde{r}_i^2 + \boldsymbol{a}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{a} + \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i\right\} + \phi^{-1} \tilde{r}_i \boldsymbol{z}_i^\top \boldsymbol{a} - 2^{-1} \ln(2\pi\phi)$. The form for $\underline{\ell}_{\text{Norm}}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ follows by recognizing $\tilde{r}_i^2 + \boldsymbol{a}^\top \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{a} - 2\tilde{r}_i \boldsymbol{z}_i^\top \boldsymbol{a} = \left(\tilde{r}_i - \boldsymbol{z}_i^\top \boldsymbol{a}\right)^2$.

A.2 Score equations for Updating Coefficients in the Variational Approximations Approach

For the Poisson response case in Section 3.1 of the main text, we update $(\boldsymbol{\kappa}, \boldsymbol{a})$ by fitting a log-link Poisson GLM with linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}$, an offset equal to $2^{-1} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i$, and a quadratic penalty of $2^{-1} \sum_{j=1}^{q} \lambda_j \boldsymbol{a}_j^\top \boldsymbol{S}_j \boldsymbol{a}_j$. The relevant score equations are then

$$\frac{\partial \underline{\ell}_{\text{Pois}}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{\kappa}} = \sum_{i=1}^{n} \left\{ y_i - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} + \frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i \right) \right\} \boldsymbol{x}_i$$
$$\frac{\partial \underline{\ell}_{\text{Pois}}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{a}} = \sum_{i=1}^{n} \left\{ y_i - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} + \frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i \right) \right\} \boldsymbol{z}_i - \sum_{j=1}^{q} \lambda_j \boldsymbol{S}_j \boldsymbol{a}_j.$$

For the normal response case in Section 3.2 of the main text, we update $(\boldsymbol{\kappa}, \boldsymbol{a})$ by fitting a linear model with linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}$, and a quadratic penalty of $2^{-1} \sum_{j=1}^{q} \lambda_j \boldsymbol{a}_j^\top \boldsymbol{S}_j \boldsymbol{a}_j$. The relevant

score equations are then

$$\frac{\partial \underline{\ell}_{\text{Norm}}(\boldsymbol{\Psi},\boldsymbol{\xi})}{\partial \boldsymbol{\kappa}} = \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\kappa} - \boldsymbol{z}_i^\top\boldsymbol{a}\right)\boldsymbol{x}_i$$
$$\frac{\partial \underline{\ell}_{\text{Norm}}(\boldsymbol{\Psi},\boldsymbol{\xi})}{\partial \boldsymbol{a}} = \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\kappa} - \boldsymbol{z}_i^\top\boldsymbol{a}\right)\boldsymbol{z}_i - \sum_{j=1}^{q}\lambda_j\boldsymbol{S}_j\boldsymbol{a}_j.$$

Conditional on the other parameters, this leads to closed-form updates $\boldsymbol{\kappa} = \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^\top\boldsymbol{a})\boldsymbol{x}_i$ and $\boldsymbol{a} = \left(\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^\top + \phi\boldsymbol{S}_{\boldsymbol{\lambda}}\right)^{-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\kappa})\boldsymbol{z}_i$.

Finally, for the Bernoulli response case in Section 3.3, we update $(\boldsymbol{\kappa},\boldsymbol{a})$ by fitting a logistic regression model with linear predictor $\boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{a}$, an offset equal to $2^{-1}\boldsymbol{z}_i^\top\boldsymbol{A}\boldsymbol{z}_i$, and a quadratic penalty of $2^{-1}\sum_{j=1}^{q}\lambda_j\boldsymbol{a}_j^\top\boldsymbol{S}_j\boldsymbol{a}_j$. The relevant score equations are then

$$\frac{\partial \underline{\ell}_{\text{Bern}}(\boldsymbol{\Psi},\boldsymbol{\xi})}{\partial \boldsymbol{\kappa}} = \sum_{i=1}^{n}(y_i - \mu_i)\,\boldsymbol{x}_i$$
$$\frac{\partial \underline{\ell}_{\text{Bern}}(\boldsymbol{\Psi},\boldsymbol{\xi})}{\partial \boldsymbol{a}} = \sum_{i=1}^{n}(y_i - \mu_i)\,\boldsymbol{z}_i - \sum_{j=1}^{q}\lambda_j\boldsymbol{S}_j\boldsymbol{a}_j,$$

where $\mu_i = \left\{1 + \exp\left(\boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{a} + \frac{1}{2}\boldsymbol{z}_i^\top\boldsymbol{A}\boldsymbol{z}_i\right)\right\}^{-1}\exp\left(\boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{a} + \frac{1}{2}\boldsymbol{z}_i^\top\boldsymbol{A}\boldsymbol{z}_i\right)$. In both the Poisson and Bernoulli response case, the estimates can be obtained by using a penalized iterative reweighted least-squares approach, as detailed in Sections 4.2 and 4.3 of Wood (2006).

A.3 Derivation of $\boldsymbol{I}_{\mathbf{v}}(\hat{\boldsymbol{\Psi}},\hat{\boldsymbol{\lambda}})$ for Common Responses

For Poisson distributed responses, let $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{\beta}$. Then ignoring constants, we have $\ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i\eta_i - \exp(\eta_i)\} + \sum_{j=1}^{q}\left(2^{-1}d_j\ln(\lambda_j) - 2^{-1}\lambda_j\boldsymbol{\beta}_j^\top\boldsymbol{S}_j\boldsymbol{\beta}_j\right)$. Letting $r_i = y_i - \exp(\eta_i)$, we obtain

$$\frac{\partial \ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})} = \left(\sum_{i=1}^{n}r_i\boldsymbol{x}_i, \frac{d_1}{2\lambda_1} - \frac{\boldsymbol{\beta}_1^\top\boldsymbol{S}_1\boldsymbol{\beta}_1}{2}, \ldots, \frac{d_q}{2\lambda_q} - \frac{\boldsymbol{\beta}_q^\top\boldsymbol{S}_q\boldsymbol{\beta}_q}{2}\right)$$

$$-\frac{\partial^2 \ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})^\top} = \begin{pmatrix} \sum_{i=1}^{n} \exp(\eta_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top & 0 & \dots & 0 \\ 0 & \frac{d_1}{2\lambda_1^2} & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{d_q}{2\lambda_q^2} \end{pmatrix}.$$

For normally distributed responses, let $r_i = y_i - \boldsymbol{x}_i^\top\boldsymbol{\kappa} - \boldsymbol{z}_i^\top\boldsymbol{\beta}$. Then ignoring constants, we have $\ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta}) = -2^{-1}n\ln(\phi) - (2\phi)^{-1}\sum_{i=1}^{n} r_i^2 + \sum_{j=1}^{q}\left(2^{-1}d_j\ln(\lambda_j) - 2^{-1}\lambda_j\boldsymbol{\beta}_j^\top\boldsymbol{S}_j\boldsymbol{\beta}_j\right)$, from which we can obtain

$$\frac{\partial \ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})} = \left(\sum_{i=1}^{n}\frac{r_i\boldsymbol{x}_i}{\phi}, -\frac{n}{2\phi} + \sum_{i=1}^{n}\frac{r_i^2}{2\phi^2}, \frac{d_1}{2\lambda_1} - \frac{\boldsymbol{\beta}_1^\top\boldsymbol{S}_1\boldsymbol{\beta}_1}{2}, \dots, \frac{d_q}{2\lambda_q} - \frac{\boldsymbol{\beta}_q^\top\boldsymbol{S}_q\boldsymbol{\beta}_q}{2}\right)$$

$$-\frac{\partial^2 \ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})^\top} = \begin{pmatrix} \sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\phi} & \sum_{i=1}^{n}\frac{\boldsymbol{x}_i r_i}{\phi^2} & 0 & \dots & 0 \\ \sum_{i=1}^{n}\frac{r_i\boldsymbol{x}_i^\top}{\phi} & -\frac{n}{2\phi^2} + \sum_{i=1}^{n}\frac{r_i^2}{\phi^3} & 0 & \dots & 0 \\ 0 & 0 & \frac{d_1}{2\lambda_1^2} & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \frac{d_q}{2\lambda_q^2} \end{pmatrix}.$$

Finally, for Bernoulli responses, we can write the complete log-likelihood as $\ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i\eta_i - \ln\{1+\exp(\eta_i)\}\right] + \sum_{j=1}^{q}\left(2^{-1}d_j\ln(\lambda_j) - 2^{-1}\lambda_j\boldsymbol{\beta}_j^\top\boldsymbol{S}_j\boldsymbol{\beta}_j\right)$, where $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{\beta}$ and quantities constant as a function of $\boldsymbol{\Psi}$ and $\boldsymbol{\beta}$ are omitted. From straightforward differentiation, we obtain

$$\frac{\partial \ell_{\text{com}}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})} = \left(\sum_{i=1}^{n} r_i\boldsymbol{x}_i, \frac{d_1}{2\lambda_1} - \frac{\boldsymbol{\beta}_1^\top\boldsymbol{S}_1\boldsymbol{\beta}_1}{2}, \dots, \frac{d_q}{2\lambda_q} - \frac{\boldsymbol{\beta}_q^\top\boldsymbol{S}_q\boldsymbol{\beta}_q}{2}\right)$$

$$-\frac{\partial^2 \ell_{\text{com},i}(\boldsymbol{\Psi},\boldsymbol{\beta})}{\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})\partial(\boldsymbol{\Psi},\boldsymbol{\lambda})^\top} = \begin{pmatrix} \sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top & 0 & \dots & 0 \\ 0 & \frac{d_1}{2\lambda_1^2} & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{d_q}{2\lambda_q^2} \end{pmatrix},$$

where $r_i = y_i - \exp(\eta_i)\{1+\exp(\eta_i)\}^{-1}$ and $w_i = \exp(\eta_i)\{1+\exp(\eta_i)\}^{-2}$.

A.4 Proof of Lemma 1

For all of the developments below, it is important to point out that notation-wise, $\boldsymbol{\beta}$ and $\boldsymbol{a}$ will be used interchangeably to denote the smoothing coefficients. Specifically, while the true parameters are denoted by $\boldsymbol{\beta}^0$ and the variational estimates are denoted by $\hat{\boldsymbol{a}}$, general reference to the smoothing coefficients will be made as $\boldsymbol{\theta} = (\boldsymbol{\Psi}^\top, \boldsymbol{\beta}^\top)^\top$ or $\boldsymbol{\theta} = (\boldsymbol{\Psi}^\top, \boldsymbol{a}^\top)^\top$.

We will prove the results separately for each of the three response types.

Normal response: We have $\boldsymbol{A} = \left(\boldsymbol{S}_{\boldsymbol{\lambda}} + \phi^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1} = n^{-1}\left(n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} + n^{-1}\phi^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1}$. Since $\lambda_j = o(n^{1/2})$, then $n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} = o(1)$ element-wise and we need only focus on the second term in the denominator. Next, let $\boldsymbol{B}_i = (\boldsymbol{x}_i^\top, \boldsymbol{z}_i^\top)^\top$ be the full $(p+d)$-vector of covariates for observation i. Applying Condition (C4) to the normal response case, we have that $\mathscr{I}(\boldsymbol{\theta}^0) = (\phi^0)^{-1}\boldsymbol{B}_1\boldsymbol{B}_1^\top$ is a finite and positive definite and hence the principled submatrix $\mathscr{I}_{\boldsymbol{z}}(\boldsymbol{\theta}^0) = (\phi^0)^{-1}\boldsymbol{z}_1\boldsymbol{z}_1^\top$ is also finite and positive definite. Since ϕ is chosen to satisfy $\phi \to \phi^0$, then by independence of the observations it follows that $n^{-1}\phi^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i^\top$ converges to $\mathscr{I}_{\boldsymbol{z}}(\boldsymbol{\theta}^0)$. It follows that $\boldsymbol{A} = O_p(n^{-1})$ element-wise.

Poisson response: We write $\boldsymbol{A}^{(1)} = \left(\boldsymbol{S}_{\boldsymbol{\lambda}} + \sum_{i=1}^{n} w_i \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1} = n^{-1}\left(n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} + n^{-1}\sum_{i=1}^{n} w_i \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1}$ where $w_i = \exp\left(\boldsymbol{x}_i^\top\boldsymbol{\kappa} + \boldsymbol{z}_i^\top\boldsymbol{a} + 2^{-1}\boldsymbol{z}_i^\top\boldsymbol{A}^{(0)}\boldsymbol{z}_i\right)$. Since $n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} = o(1)$ element-wise, then we can solely focus on the second term in the denominator. Applying Condition (C4), we obtain $\mathscr{I}(\boldsymbol{\theta}^0) = w_1^0\boldsymbol{B}_1\boldsymbol{B}_1^\top$ is finite and positive definite where $w_1^0 = \exp\left(\boldsymbol{x}_1^\top\boldsymbol{\kappa}^0 + \boldsymbol{z}_1^\top\boldsymbol{\beta}^0\right)$, and hence $\mathscr{I}_{\boldsymbol{z}}(\boldsymbol{\theta}^0) =$

$w_1^0 \boldsymbol{z}_1 \boldsymbol{z}_1^\top$ is also finite and positive definite. Suppose we take $\boldsymbol{A}^{(0)} = n^{-r}\boldsymbol{I}_d$ for $r > 2^{-1}$ as an starting value, where $\boldsymbol{I}_d$ denotes an identity matrix of dimension d. Since $\boldsymbol{\theta}$ is chosen to satisfy $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| = O(d^{1/2}n^{-1/2})$, then by independence of the observations it follows that $n^{-1}\sum_{i=1}^{n} w_i \boldsymbol{z}_i \boldsymbol{z}_i^\top$ converges to $\mathscr{I}_{\boldsymbol{z}}(\boldsymbol{\theta}^0)$. It follows that $\boldsymbol{A} = O_p(n^{-1})$ element-wise.

Bernoulli response: Write $\boldsymbol{A}^{(1)} = \left(\boldsymbol{S}_{\boldsymbol{\lambda}} + \sum_{i=1}^{n} \tilde{w}_i \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1} = n^{-1}\left(n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} + n^{-1}\sum_{i=1}^{n} \tilde{w}_i \boldsymbol{z}_i \boldsymbol{z}_i^\top\right)^{-1}$ where $\tilde{w}_i = \left\{1 + \exp\left(-\boldsymbol{x}_i^\top \boldsymbol{\kappa} - \boldsymbol{z}_i^\top \boldsymbol{a} - 2^{-1}\boldsymbol{z}_i^\top \boldsymbol{A}^{(0)} \boldsymbol{z}_i\right)\right\}^{-1}$. Since $\lambda_j = o(n^{1/2})$, then $n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}} = o(1)$ element-wise and we need only focus on the second term in the denominator. Suppose we take $\boldsymbol{A}^{(0)} = n^{-r}\boldsymbol{I}_d$ for $r > 2^{-1}$ as a starting value, where $\boldsymbol{I}_d$ denotes an identity matrix of dimension d. Since $\boldsymbol{\theta}$ is chosen to satisfy $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| = O(d^{1/2}n^{-1/2})$, then by independence of the observations it follows that $n^{-1}\sum_{i=1}^{n} \tilde{w}_i \boldsymbol{z}_i \boldsymbol{z}_i^\top$ converges to $\tilde{\mathscr{I}}(\boldsymbol{\theta}^0) = \tilde{w}_1^0 \boldsymbol{z}_1 \boldsymbol{z}_1^\top$ where $\tilde{w}_1^0 = \left\{1 + \exp\left(-\boldsymbol{x}_1^\top \boldsymbol{\kappa}^0 - \boldsymbol{z}_1^\top \boldsymbol{\beta}^0\right)\right\}^{-1}$. It remains then to prove that this matrix is finite and positive definite. To show this, note that $1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0\right) > 1$ and hence it is straightforward to show $\tilde{w}_1^0 > w_1^0$ where $w_1^0 = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0\right)\left\{1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0\right)\right\}^{-2}$. Next, applying Condition (C4) we have that $\mathscr{I}(\boldsymbol{\theta}^0) = w_1^0 \boldsymbol{B}_1 \boldsymbol{B}_1^\top$ is finite and positive definite, and hence $\mathscr{I}_{\boldsymbol{z}}(\boldsymbol{\theta}^0) = w_1^0 \boldsymbol{z}_1 \boldsymbol{z}_1^\top$ is also finite and positive definite. Since $\tilde{w}_1^0 > w_1^0$, and noting that $\tilde{w}_1^0$ is bounded by some sufficiently large constant under Condition (C3), then $\tilde{\mathscr{I}}(\boldsymbol{\theta}^0)$ must also be finite and positive finite. It follows that $\boldsymbol{A} = O_p(n^{-1})$ element-wise.

A.5 Proof of Theorem 1

We first prove the following result relating the variational and true model log-likelihood functions.

Proposition 1. *Assume $\boldsymbol{A} = O(n^{-1})$ element-wise, and let $R = 2^{-1}\sum_{j=1}^{q} d_j \ln(\lambda_j) - tr(\boldsymbol{S}_{\boldsymbol{\lambda}}\boldsymbol{A}) + 2^{-1}\ln\det(\boldsymbol{A})$. Then for normal responses we have*

$$\underline{\ell}_{Norm}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} - \frac{1}{2\phi}\sum_{i=1}^{n} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i + R,$$

where $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta})$ *and* $f(y_i|\boldsymbol{\theta})$ *is the normal distribution with mean* $\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{\beta}$ *and variance* ϕ.

For Poisson responses, we have

$$\underline{\ell}_{Pois}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n}\right) \right| - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R.$$

where $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta})$ *and* $f(y_i|\boldsymbol{\theta})$ *is the Poisson distribution with mean* $\exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{\beta})$.

For Bernoulli responses, we have

$$\underline{\ell}_{Bern}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n}\right) \right| - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R,$$

where $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta})$ *and* $f(y_i|\boldsymbol{\theta})$ *is the Bernoulli distribution with mean* $\{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\kappa} - \boldsymbol{z}_i^\top \boldsymbol{\beta})\}^{-1}$.

Proof. The normal response case is trivial by realizing that $\ell(\boldsymbol{\theta}) = -2^{-1} n \ln(\phi) - (2\phi)^{-1} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\kappa} - \boldsymbol{z}_i^\top \boldsymbol{\beta}\right)^2$.

For both the Poisson and Bernoulli response case we will use the fact that under Conditions (C2) and (C6) and given $\boldsymbol{A} = O(n^{-1})$ element-wise, we obtain $\boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i \leq \|\boldsymbol{z}_i^\top\|^2 \|\boldsymbol{A}\| = O(d^2 n^{-1}) = o(1)$ for all $i = 1, \ldots, n$.

Turning specifically to the Poisson response case, by using the Taylor expansion $\exp(2^{-1} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i) = 1 + O(\boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i)$ for $\boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i \to 0$, we can write the variational log-likelihood as

$$\begin{aligned}
\underline{\ell}_{\text{Pois}}(\boldsymbol{\Psi}, \boldsymbol{\xi}) &= \sum_{i=1}^{n} \left\{ y_i \left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) \exp\left(\frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i \right) \right\} - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R \\
&= \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) \left\{ 1 + O\left(\frac{d^2}{n} \right) \right\} \right] - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R \\
&= \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n}\right) \right| - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R,
\end{aligned}$$

where the constant $-\sum_{i=1}^{n} \ln(y_i!)$ is added in the last step.

For the Bernoulli response case, we will use the fact that $\ln(c+\varepsilon) = \ln(c) + O(\varepsilon)$ if $\varepsilon \to 0$ and c is positive and bounded away from zero. Choosing $\varepsilon = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}\right)\left(\exp\left(2^{-1}\boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i\right) - 1\right)$, and recognizing that this satisfies $\varepsilon \to 0$ given Conditions (C2) and (C6) and $\boldsymbol{A} = O(n^{-1})$ element-wise, then we can write the variational log-likelihood as follows

$$
\begin{aligned}
\underline{\ell}_{\text{Bern}}(\boldsymbol{\Psi}, \boldsymbol{\xi}) &= \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) - \ln \left\{ 1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} + \frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i \right) \right\} \right] + R \\
&= \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) \right. \\
&\quad \left. - \ln \left\{ 1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a} \right) \left(\exp\left(\frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i \right) - 1 \right) \right\} \right] + R \\
&= \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \left| O\left\{ \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left(\exp\left(\frac{1}{2} \boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i \right) - 1 \right) \right\} \right| - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R \\
&= \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n} \right) \right| - \frac{1}{2} \boldsymbol{a}^\top \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} + R,
\end{aligned}
$$

where to go from the third to the fourth expressions on the right hand side, we again use the result $\exp(2^{-1}\boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i) = 1 + O(\boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i)$ for $\boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i \to 0$. □

Proposition 1 is critical, as it allows us to then relate the first and second derivatives of the variational log-likelihood to the corresponding derivatives of true model log-likelihood as follows. To see this, write the variational log-likelihood as $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \underline{\ell}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$, and let $\nabla_{\boldsymbol{\theta}}$ and $\nabla^2_{\boldsymbol{\theta}}$ with the first and second derivative operators with respect to $\boldsymbol{\theta}$. Also, define $\boldsymbol{B}_i = (\boldsymbol{x}_i^\top, \boldsymbol{z}_i^\top)^\top$. Then we have

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) - \left(\boldsymbol{0}, -\frac{1}{2\phi^2} \sum_{i=1}^{n} \boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i, \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a} \right) \\
\nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Pois}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n} \right) \right| \boldsymbol{B}_i - (\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a}) \qquad (1) \\
\nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Bern}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) \left| O\left(\frac{d^2}{n} \right) \right| \boldsymbol{B}_i - (\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{a}),
\end{aligned}
$$

Also,

$$
\begin{aligned}
-\nabla^2_{\boldsymbol{\theta}} \underline{\ell}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= -\nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) + \boldsymbol{M}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} \\
-\nabla^2_{\boldsymbol{\theta}} \underline{\ell}_{\text{Pois}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= -\nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) + \boldsymbol{M}_{\text{Pois}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} \\
-\nabla^2_{\boldsymbol{\theta}} \underline{\ell}_{\text{Bern}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} &= -\nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) + \boldsymbol{M}_{\text{Bern}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\},
\end{aligned} \tag{2}
$$

where $\boldsymbol{M}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$ is a $(p+d+1) \times (p+d+1)$ matrix with structure

$$
\boldsymbol{M}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (4\phi^3)^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_{\boldsymbol{\lambda}} \end{pmatrix},
$$

and both $\boldsymbol{M}_{\text{Pois}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$ and $\boldsymbol{M}_{\text{Bern}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$ are $(p+d) \times (p+d)$ matrices with the following structure

$$
\begin{pmatrix} \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) | O(d^2 n^{-1}) | \boldsymbol{x}_i \boldsymbol{x}_i^\top & \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) | O(d^2 n^{-1}) | \boldsymbol{x}_i \boldsymbol{z}_i^\top \\ \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) | O(d^2 n^{-1}) | \boldsymbol{z}_i \boldsymbol{x}_i^\top & \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa} + \boldsymbol{z}_i^\top \boldsymbol{a}) | O(d^2 n^{-1}) | \boldsymbol{z}_i \boldsymbol{z}_i^\top + \boldsymbol{S}_{\boldsymbol{\lambda}} \end{pmatrix}.
$$

We now move on to the main proof for consistency. To show this, if we can prove that for any given $\varepsilon > 0$, there exists a sufficiently large constant $K_1 > 0$ such that

$$
\text{P}\left(\sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| = d^{1/2} n^{-1/2} K_1} \underline{\ell}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} < \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\} \right) \geq 1 - \varepsilon, \tag{3}
$$

there it implies that with probability tending to 1 there exists a local maximizer of $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ satisfying $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_p(d^{1/2} n^{-1/2})$.

To prove (3), first note that any point inside the ball $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| = d^{1/2} n^{-1/2} K_1\}$ satisfies the conditions of Lemma 1 and hence $\boldsymbol{A} = O(n^{-1})$ element-wise. Next applying a Taylor expansion

we have

$$\begin{aligned}\Delta(\boldsymbol{\theta}) &= \underline{\ell}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\} - \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\} \\ &= (\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top \nabla_{\boldsymbol{\theta}} \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\} - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top \left[-\nabla^2_{\boldsymbol{\theta}} \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\}\right](\boldsymbol{\theta}-\boldsymbol{\theta}^0) \\ &\quad + \frac{1}{6}\sum_{r,s,t=1}^{\dim(\theta)} \frac{\partial^3 \underline{\ell}\left\{\bar{\boldsymbol{\theta}}, \text{vech}(\boldsymbol{A})\right\}}{\partial\theta_r \partial\theta_s \partial\theta_t}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)_r(\boldsymbol{\theta}-\boldsymbol{\theta}^0)_s(\boldsymbol{\theta}-\boldsymbol{\theta}^0)_t \\ &\triangleq T_1 + T_2 + T_3,\end{aligned}$$

where $\bar{\boldsymbol{\theta}}$ lies on the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^0$ and $\dim(\boldsymbol{\theta}) = p+d+1$ for normal responses and $p+d$ for Poisson and Bernoulli responses.

For term T_1, by the Cauchy-Schwarz inequality we obtain $T_1 \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \, \|\nabla_{\boldsymbol{\theta}} \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\}\| = d^{1/2} K_1 \|n^{-1/2} \nabla_{\boldsymbol{\theta}} \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\}\|$. Next, we establish the following three results. First by Conditions (C2)-(C3) and (C6), Lemma 1, and applying the Cauchy-Schwarz inequality, it holds that $n^{-1/2}\left|\sum_{i=1}^n \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i\right| \leq n^{-1/2}\sum_{i=1}^n \|\boldsymbol{z}_i\|^2 \|\boldsymbol{A}\| = O(d^2 n^{-1/2}) = o(1)$ and hence $2^{-1} n^{-1/2} (\phi^0)^{-2}\left|\sum_{i=1}^n \boldsymbol{z}_i^\top \boldsymbol{A} \boldsymbol{z}_i\right| = o(1)$. Second, $n^{-1/2}\left\|\boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{\beta}^0\right\| \leq o(d^{1/2})$ by Conditions (C3), (C6) and $\lambda_j = o(n^{1/2})$ for all $j = 1, \ldots, q$. Third, $n^{-1/2}\left\|\sum_{i=1}^n \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0) |O(d^2 n^{-1})| \boldsymbol{B}_i\right\| = O(d^{5/2} n^{-1/2}) = o(d^{1/2})$ by Conditions (C2)-(C3) and (C6).

Applying the above set of three results, we obtain $\|n^{-1/2} \nabla_{\boldsymbol{\theta}} \underline{\ell}\left\{\boldsymbol{\theta}^0, \text{vech}(\boldsymbol{A})\right\}\| \leq \left\|n^{-1/2} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^0)\right\| + o(d^{1/2})$. Finally, since $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^0) = \sum_{i=1}^n \partial \ln\{f(y_i|\boldsymbol{\theta}^0)\}/\partial\boldsymbol{\theta}$ where $f(y_i|\boldsymbol{\theta})$ belongs to the exponential family for all three responses considered, then under Conditions (C1)-(C4) we can utilize standard asymptotic developments regarding generalized linear models to show that $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^0) = O_p(n^{1/2})$ element-wise (e.g., Zou, 2006). We thus obtain $T_1 = K_1 O_p(d)$.

For term T_2, using (2) we can write $T_2 = -2^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\{-\nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0) - 2^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top \boldsymbol{M}_{\text{resp}}(\boldsymbol{\theta}^{\boldsymbol{0}})(\boldsymbol{\theta}-\boldsymbol{\theta}^0)$ where $\boldsymbol{M}_{\text{resp}}\{\boldsymbol{\theta}^{\boldsymbol{0}}, \text{vech}(\boldsymbol{A})\}$ refers to one of $\boldsymbol{M}_{\text{Norm}}\{\boldsymbol{\theta}^{\boldsymbol{0}}, \text{vech}(\boldsymbol{A})\}$, $\boldsymbol{M}_{\text{Pois}}\{\boldsymbol{\theta}^{\boldsymbol{0}}, \text{vech}(\boldsymbol{A})\}$, or $\boldsymbol{M}_{\text{Bern}}\{\boldsymbol{\theta}^{\boldsymbol{0}}, \text{vech}(\boldsymbol{A})\}$ depending on the response type. By the Cauchy-

Schwarz inequality,

$$(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top \boldsymbol{M}_{\text{resp}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0) \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|^2 \, \|\boldsymbol{M}_{\text{resp}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}\|$$
$$= dK_1^2 \|n^{-1}\boldsymbol{M}_{\text{resp}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}\|.$$

Now, for the case of normal responses, we have $n^{-1}\left|\sum_{i=1}^{n} \boldsymbol{z}_i^\top \boldsymbol{A}\boldsymbol{z}_i\right| = o(1)$ by Condition (C2) and Lemma 1. Also, $\|n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}}\| = O(\lambda d n^{-1}) = o(1)$ by Condition (C6), and hence we conclude $\|n^{-1}\boldsymbol{M}_{\text{Norm}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}\| = o(1)$. Turning to Poisson and Bernoulli responses, given $\|n^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}}\| = o(1)$ then we need only focus on the the matrix $n^{-1}\sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0)|O(d^2 n^{-1})|\boldsymbol{B}_i \boldsymbol{B}_i^\top$. By Conditions (C3), it holds that $\exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0) < K_2 < \infty$ for all $i = 1, \ldots, n$ and some constant K_2. Thus by Conditions (C2) and (C6), we have

$$\left\| n^{-1}\sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\kappa}^0 + \boldsymbol{z}_i^\top \boldsymbol{\beta}^0)|O(d^2 n^{-1})|\boldsymbol{B}_i \boldsymbol{B}_i^\top \right\| = O\left(\frac{d^2}{n}\right) \left\| n^{-1}\sum_{i=1}^{n} \boldsymbol{B}_i \boldsymbol{B}_i^\top \right\|$$
$$= O\left(\frac{d^3}{n}\right) = o(1)$$

It follows that $\left\|n^{-1}\boldsymbol{M}_{\text{Pois}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}\right\| = o(1)$ and $\left\|n^{-1}\boldsymbol{M}_{\text{Bern}}\{\boldsymbol{\theta}^{\mathbf{0}}, \text{vech}(\boldsymbol{A})\}\right\| = o(1)$, and hence $T_2 = -2^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\{-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0) + K_1^2 o(d)$.

Write the first term of T_2 and $-2^{-1}n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\{-n^{-1}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)$. Then we make use of the following result.

Proposition 2. $\left\|-n^{-1}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) - \mathscr{I}(\boldsymbol{\theta}^0)\right\| = O_p\left(d^{-1}\right)$

Proof. By Markov's inequality and independence of the observations,

$$\mathrm{P}\left(\left\|-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) - \mathscr{I}(\boldsymbol{\theta}^0)\right\| \geq \frac{\varepsilon_2}{d}\right) \leq \frac{d^2}{n^2\varepsilon_2^2}\sum_{i=1}^{n} \mathrm{E}\left(\sum_{r,s=1}^{\dim(\boldsymbol{\theta})}\left\{\frac{\partial^2 \ln\{f(y_i|\boldsymbol{\theta}^0)\}}{\partial\theta_r\partial\theta_s} - \mathrm{E}\left(\frac{\partial^2 \ln\{f(y_i|\boldsymbol{\theta}^0)\}}{\partial\theta_r\partial\theta_s}\right)\right\}^2\right)$$
$$= \frac{d^2}{n^2\varepsilon_2^2}\, nO(d^2) \quad \text{by Condition (C5)}$$
$$= O\left(\frac{d^4}{n}\right) = o(1) \quad \text{by Condition (C6).}$$

and the required result follows. □

Applying the above proposition, we obtain

$$-\frac{1}{2}n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\{-n^{-1}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0) = -\frac{1}{2}n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\left\{\mathscr{I}(\boldsymbol{\theta}^0)+O_p\left(\frac{1}{d}\right)\right\}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)$$
$$= -\frac{1}{2}n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\mathscr{I}(\boldsymbol{\theta}^0)(\boldsymbol{\theta}-\boldsymbol{\theta}^0)+O_p(1)$$

Therefore, $T_2 = -2^{-1}n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\mathscr{I}(\boldsymbol{\theta}^0)(\boldsymbol{\theta}-\boldsymbol{\theta}^0)+K_1^2 o(d)$. Finally, by Condition (C4) we have the minimum eigenvalue of $\mathscr{I}(\boldsymbol{\theta}^0)$, denoted here as $\tau_{\min}$, is bounded away from zero. Thus $n(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\mathscr{I}(\boldsymbol{\theta}^0)(\boldsymbol{\theta}-\boldsymbol{\theta}^0) \geq n\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|^2\tau_{\min} = K_1^2\tau_{\min}d$ and hence $T_2 \leq -K_1^2\tau_{\min}d < 0$.

For term T_3, by the Cauchy-Schwarz and Minowski inequalities, we can write

$$6|T_3| \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|^3\left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})}\left(\frac{\partial^3\underline{\ell}\left\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\right\}}{\partial\theta_r\partial\theta_s\partial\theta_t}\right)^2\right)^{1/2}$$
$$\leq \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|^3\left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})}\left(\frac{\partial^3\ell(\bar{\boldsymbol{\theta}})}{\partial\theta_r\partial\theta_s\partial\theta_t}\right)^2\right)^{1/2} + \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|^3\left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})}\left(\frac{\partial[\boldsymbol{M}_{\mathrm{resp}}\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\}]_{rs}\}}{\partial\theta_t}\right)^2\right)^{1/2}$$
$$\triangleq U_1+U_2$$

where $[\boldsymbol{M}\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\}]_{rs}$ refers to element (r,s) for $\boldsymbol{M}\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\}$. Dealing with term U_2 first, in the normal response case, from equation (2) we have $\partial[\boldsymbol{M}_{\mathrm{Norm}}\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\}]_{rs}\}/\partial\theta_t = 0$ for all $r,s,t = 1,\ldots,\dim(\boldsymbol{\theta})$ except for $\partial[\boldsymbol{M}_{\mathrm{Norm}}\{\bar{\boldsymbol{\theta}},\mathrm{vech}(\boldsymbol{A})\}]_{p+1,p+1}\}/\partial\phi = -12^{-1}\bar{\phi}^{-4}\sum_{i=1}^n \boldsymbol{z}_i^\top\boldsymbol{A}\boldsymbol{z}_i$. Therefore since $\bar{\phi}$ lies inside the ball $\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| = d^{1/2}n^{-1/2}K_1\}$, then we obtain

$$U_2 = -12K_1^3\left(\frac{d}{n}\right)^{3/2}\frac{1}{\bar{\phi}^4}\left|\sum_{i=1}^n\boldsymbol{z}_i^\top\boldsymbol{A}\boldsymbol{z}_i\right| \leq -12K_1^3\left(\frac{d}{n}\right)^{3/2}\frac{1}{\bar{\phi}^4}\sum_{i=1}^n\|\boldsymbol{z}_i\|^2\|\boldsymbol{A}\|$$
$$= K_1^3 O_p\left(\frac{d^{7/2}}{n^{3/2}}\right) \quad \text{by Conditions (C2)-(C3) and Lemma 1}$$
$$= o_p(1) \quad \text{by Condition (C6).}$$

For Poisson response case, from equation (2), and denoting $\boldsymbol{B}_i = (\boldsymbol{x}_i^\top, \boldsymbol{z}_i^\top)^\top$, we have $\partial[\boldsymbol{M}_{\text{Pois}}\{\bar{\boldsymbol{\theta}}, \text{vech}(\boldsymbol{A})\}]_{rs}\}/\partial\theta_t = \sum_{i=1}^{n} \exp\left(\boldsymbol{x}_i^\top \bar{\boldsymbol{\kappa}} + \boldsymbol{z}_i^\top \bar{\boldsymbol{\beta}}\right) \left|O_p(d^2 n^{-1})\right| \boldsymbol{B}_{ir}\boldsymbol{B}_{is}\boldsymbol{B}_{it}$ for $r,s,t = 1,\ldots,\dim(\boldsymbol{\theta})$ where $\boldsymbol{B}_{ir}$ refers to the r-th element of $\boldsymbol{B}_i$. The same expression in obtained by the Bernoulli response case. Since $(\bar{\boldsymbol{\kappa}}, \bar{\boldsymbol{\beta}})$ lies inside the ball $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| = d^{1/2} n^{-1/2} K_1\}$, then we obtain

$$
\begin{aligned}
U_2 &= K_1^3 \left(\frac{d}{n}\right)^{3/2} \left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})} \left(\sum_{i=1}^{n} \exp\left(\boldsymbol{x}_i^\top \bar{\boldsymbol{\kappa}} + \boldsymbol{z}_i^\top \bar{\boldsymbol{\beta}}\right) \left|O_p(d^2 n^{-1})\right| \boldsymbol{B}_{ir}\boldsymbol{B}_{is}\boldsymbol{B}_{it} \right)^2 \right)^{1/2} \\
&= K_1^3 \left(\frac{d}{n}\right)^{3/2} O_p\left(d^{7/2}\right) \quad \text{by Conditions (C2)-(C3)} \\
&= o_p(1) \quad \text{by Condition (C6).}
\end{aligned}
$$

It remains then to determine the order of term U_1. We have

$$
\begin{aligned}
U_1 &= K_1^3 \left(\frac{d}{n}\right)^{3/2} \left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})} \left(\frac{\partial^3 \ell(\bar{\boldsymbol{\theta}})}{\partial\theta_r \partial\theta_s \partial\theta_t} \right)^2 \right)^{1/2} \\
&= K_1^3 \left(\frac{d}{n}\right)^{3/2} n\, O_p \left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})} \left(\frac{\partial^3 \ln f(y_1|\bar{\boldsymbol{\theta}})}{\partial\theta_r \partial\theta_s \partial\theta_t} \right)^2 \right)^{1/2} \quad \text{by independence of the observations} \\
&\leq \frac{K_1^3 d^{3/2}}{n^{1/2}} O_p \left(\sum_{r,s,t=1}^{\dim(\boldsymbol{\theta})} G_{rst}^2(y_1|\bar{\boldsymbol{\theta}}) \right)^{1/2} \quad \text{by Condition (C5)} \\
&= \frac{K_1^3 d^{3/2}}{n^{1/2}} O_p(d^{3/2}) = K_1^3 o_p(d) \quad \text{by Condition (C6).}
\end{aligned}
$$

Combining the order results of U_1 and U_2 above, we conclude that $T_3 = o_p(d)$.

To summarize then, we have $T_1 = K_1 O_p(d), T_2 \leq -K^2 d\tau_{\min}, T_3 = o_p(d)$. It follows that if provided we choose K_1 large enough, then T_2 which is negative asymptotically dominates T_1 and T_2, and the probability statement in (3) follows. Therefore we conclude that there exists a local maximizer of $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ satisfying $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_p(d^{1/2} n^{-1/2})$. Finally, since as a function of $\boldsymbol{\theta}$ the variational log-likelihood $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ resembles a generalized linear model with a quadratic penalty on the smoothing coefficients $\boldsymbol{a}$, then it is straightforward to show that the variational log-likelihood is

strictly convex with respect to $\boldsymbol{\theta}$ for the three response types considered. Hence $\hat{\boldsymbol{\theta}}$ coincides with the VA estimate (the global maximizer) and the required result follows.

A.6 Proof of Theorem 2

Based on equation (1) in the proof of Theorem 1, we have the following expressions for the first derivatives of the variational log-likelihood evaluated at $(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\xi}}) = \{\hat{\boldsymbol{\theta}}, \text{vech}(\hat{\boldsymbol{A}})\}$,

$$\mathbf{0} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Norm}}\{\hat{\boldsymbol{\theta}}, \text{vech}(\hat{\boldsymbol{A}})\} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) - \left(\mathbf{0}, -\frac{1}{2n^{1/2}\hat{\phi}^2} \sum_{i=1}^{n} \boldsymbol{z}_i^\top \hat{\boldsymbol{A}} \boldsymbol{z}_i, \frac{1}{n^{1/2}} \boldsymbol{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{a}}\right)$$

$$\mathbf{0} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Pois}}\{\hat{\boldsymbol{\theta}}, \text{vech}(\hat{\boldsymbol{A}})\} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{n^{1/2}} \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\kappa}} + \boldsymbol{z}_i^\top \hat{\boldsymbol{a}}) \left| O\left(\frac{d^2}{n}\right) \right| \boldsymbol{B}_i - \left(\mathbf{0}, \frac{1}{n^{1/2}} \boldsymbol{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{a}}\right)$$

$$\mathbf{0} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{Bern}}\{\hat{\boldsymbol{\theta}}, \text{vech}(\hat{\boldsymbol{A}})\} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{n^{1/2}} \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\kappa}} + \boldsymbol{z}_i^\top \hat{\boldsymbol{a}}) \left| O\left(\frac{d^2}{n}\right) \right| \boldsymbol{B}_i - \left(\mathbf{0}, \frac{1}{n^{1/2}} \boldsymbol{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{a}}\right).$$

We aim to prove that for all three response types, the first derivatives of the variational log-likelihood is dominated by $\nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}})$. First, we have by Theorem 1 that $\hat{\phi} \xrightarrow{p} \phi^0$. Thus applying this result along with Conditions (C2)-(C3) and (C6') and Lemma 1, we have $n^{-1/2}\hat{\phi}^{-2} \left| \sum_{i=1}^{n} \boldsymbol{z}_i^\top \hat{\boldsymbol{A}} \boldsymbol{z}_i \right| \leq n^{-1/2}\hat{\phi}^{-2} \sum_{i=1}^{n} \|\boldsymbol{z}_i\|^2 \left\|\hat{\boldsymbol{A}}\right\| = O_p(d^2 n^{-1/2}) = o_p(d^{-1/2})$. Also, by Conditions (C6') and $\lambda_j = o(n^{1/2} d^{-1})$ for all $j = 1, \ldots, q$, we have $n^{-1/2} \boldsymbol{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{\beta}} = o(d^{-1/2})$ element-wise. Turning to the Poisson and Bernoulli response case, by the estimation consistency result from Theorem 1 along with Conditions (C2)-(C3) and (C6'), we have $n^{-1/2} \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\kappa}} + \boldsymbol{z}_i^\top \hat{\boldsymbol{\beta}}) \left| O(d^2 n^{-1}) \right| \boldsymbol{B}_i = O_p(d^2 n^{-1/2}) = o_p(d^{-1/2})$ element-wise.

Applying the above results, we therefore obtain

$$\mathbf{0} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \underline{\ell}_{\text{resp}}\{\hat{\boldsymbol{\theta}}, \text{vech}(\hat{\boldsymbol{A}})\} = \frac{1}{n^{1/2}} \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}) + \boldsymbol{\delta}_1,$$

where $\boldsymbol{\delta}_1 = o_p(d^{-1/2})$ element-wise and $\underline{\ell}_{\text{resp}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$ refers to one of $\underline{\ell}_{\text{Norm}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$, $\underline{\ell}_{\text{Pois}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$, or $\underline{\ell}_{\text{Bern}}\{\boldsymbol{\theta}, \text{vech}(\boldsymbol{A})\}$ depending on the response type. Next, applying a Taylor

series expansion to $\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}^0$, we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{n^{1/2}}\nabla_{\boldsymbol{\theta}}\ell(\hat{\boldsymbol{\theta}}) + \boldsymbol{\delta}_1 \\ &= \frac{1}{n^{1/2}}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) - \left\{-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}^0\right)\right\} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \frac{1}{2n^{1/2}}\boldsymbol{\nu} + \boldsymbol{\delta}_1\end{aligned}$$

where $\boldsymbol{\nu}$ is a vector with t-th element $\nu_t = \sum_{r,s=1}^{\dim(\theta)} \partial^3\ell(\bar{\boldsymbol{\theta}})/\partial\theta_t\partial\theta_r\partial\theta_s\ (\hat{\theta} - \theta^0)_r(\hat{\theta} - \theta^0)_s$ with $(\hat{\theta} - \theta^0)_r$ denoting the r-th element of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)$, and $\bar{\boldsymbol{\theta}}$ lies on the line segment joining $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$. Note that by the Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{n^{1/2}}|\nu_t| &\leq \frac{1}{n^{1/2}}\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathbf{0}})\|^2 \sum_{i=1}^{n}\left(\sum_{r,s=1}^{\dim(\theta)} G^2_{rst}(y_i)\right)^{1/2} \quad \text{by Condition (C5)} \\ &= O_p\left(\frac{d^2}{n^{1/2}}\right) \quad \text{by independence of the observations and Theorem 1} \\ &= o_p\left(\frac{1}{d^{1/2}}\right) \quad \text{by Condition (C6').}\end{aligned}$$

In addition, with a similar proof to that of Proposition 2, we can show that under Condition (C6'), $\left\|-n^{-1}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) - \mathscr{I}(\boldsymbol{\theta}^0)\right\| = O_p\left(d^{-3/2}\right)$. Applying these two results, we obtain

$$\mathbf{0} = \frac{1}{n^{1/2}}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) - \left\{\mathscr{I}(\boldsymbol{\theta}^0) + O_p\left(\frac{1}{d^{3/2}}\right)\right\} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \frac{1}{2n^{1/2}}\boldsymbol{\nu} + \boldsymbol{\delta}_1. \tag{4}$$

Furthermore, rearranging equation (4) we obtain $n^{1/2}\mathscr{I}(\boldsymbol{\theta}^0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = n^{-1/2}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = o_p(d^{-1/2})$ element-wise.

Let $\boldsymbol{G}$ be a $k \times \dim(\boldsymbol{\theta})$ matrix with $k < d$ a fixed constant and satisfying $\boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{I}_k$. Then following straightforward manipulation, we obtain

$$n^{1/2}\boldsymbol{G}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \frac{1}{n^{1/2}}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) + \boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{\delta}. \tag{5}$$

We now show that the last term on the right hand side of (5) is asymptotically negligible. To do this, let $\tau_{\max}(\cdot)$ and $\tau_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix. As $\mathscr{I}^{-1}(\boldsymbol{\theta}^0)$

and $\boldsymbol{G}^\top\boldsymbol{G}$ are both positive-semidefinite matrix, it follows that

$$\begin{aligned}\tau_{\max}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\} &\leq \tau_{\max}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\}^2\tau_{\max}(\boldsymbol{G}^\top\boldsymbol{G})\\ &= \tau_{\max}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\}^2\tau_{\max}(\boldsymbol{G}\boldsymbol{G}^\top)\\ &= \frac{1}{\tau_{\min}\{\mathscr{I}(\boldsymbol{\theta}^0)\}^2}\\ &= O(1) \quad \text{by Condition (C4).}\end{aligned}$$

Therefore, $\|\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{\delta}\|^2 \leq \tau_{\max}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\}\|\boldsymbol{\delta}\|^2 = o_p(1)$ since $\boldsymbol{\delta} = o_p(d^{-1/2})$ element-wise. Equation (5) hence reduces to

$$n^{1/2}\boldsymbol{G}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \frac{1}{n^{1/2}}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0) + o_p(1).$$

To establish the asymptotic normality of the VA estimates, we let $\boldsymbol{L}_i = n^{-1/2}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\partial \ln f(y_i|\boldsymbol{\theta}^0)/\partial\boldsymbol{\theta}$ such that $\sum_{i=1}^{n}\boldsymbol{L}_i = n^{-1/2}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0)$. Note that $\mathrm{E}(\boldsymbol{L}_i) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{L}_i) = n^{-1}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top$ by Conditions (C3) and (C4) respectively. We need to show the Lindeberg condition is satisfied. That is, for any $\varepsilon > 0$

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Bigg[\{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\}^\top\left(\frac{1}{n}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top\right)^{-1}\{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\}\\ \times\mathbb{1}_{\{\boldsymbol{L}_i-\mathrm{E}(\boldsymbol{L}_i)\}^\top\left(\frac{1}{n}\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top\right)^{-1}\{\boldsymbol{L}_i-\mathrm{E}(\boldsymbol{L}_i)\}>n\varepsilon}\Bigg] = 0\end{aligned}$$

To see the above is true, first note that from the prove of Theorem 1 it is straightforward to show $\boldsymbol{L}_i = O_p(n^{-1/2})$ element-wise (see also Theorem 14.4.1, Bishop et al., 2007). Therefore $\mathrm{E}\left(\|\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\|^2\right)^2 = \mathrm{E}\left(\|\boldsymbol{L}_i)\|^2\right)^2 = O_p\left(n^{-2}\right)$. By Markov's inequality then,

$$\mathrm{E}\left(\mathbb{1}_{\|\boldsymbol{L}_i-\mathrm{E}(\boldsymbol{L}_i)\|^2>\frac{\varepsilon}{d}}\right)^2 = \mathrm{P}\left(\|\boldsymbol{L}_i\|^2 \geq \frac{\varepsilon}{d}\right) \leq \frac{d}{\varepsilon}\mathrm{E}(\|\boldsymbol{L}_i\|^2) = O_p\left(\frac{d}{n}\right).$$

Next, note that $\tau_{\max}\{(\boldsymbol{G}\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top)^{-1}\} = \tau_{\min}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{G}^\top\boldsymbol{G}\}^{-1} = \tau_{\min}\{\mathscr{I}^{-1}(\boldsymbol{\theta}^0)\}^{-1} = \tau_{\max}\{\mathscr{I}(\boldsymbol{\theta}^0)\}$.

Furthermore, by the Gershgorin circle theorem $\tau_{\max}(\mathscr{I}(\boldsymbol{\theta}^0))$ is at most of order $O_p(d)$. It follows that

$$
\begin{aligned}
T &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\}^\top \left(\frac{1}{n} \boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top \right)^{-1} \{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\} \right. \\
&\quad \left. \times \mathbb{1}_{\{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\}^\top \left(\frac{1}{n} \boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top \right)^{-1} \{\boldsymbol{L}_i - \mathrm{E}(\boldsymbol{L}_i)\} > n\varepsilon} \right] \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\left(\boldsymbol{L}_i^\top \left(\boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top \right)^{-1} \boldsymbol{L}_i \times \mathbb{1}_{\boldsymbol{L}_i^\top \left(\boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top \right)^{-1} \boldsymbol{L}_i > \varepsilon} \right) \\
&\leq \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\left(\tau_{\max}\{\mathscr{I}(\boldsymbol{\theta}^0)\} \|\boldsymbol{L}_i\|^2 \times \mathbb{1}_{\|\boldsymbol{L}_i\|^2 > \varepsilon \tau_{\max}\{\mathscr{I}(\boldsymbol{\theta}^0)\}^{-1}} \right) \\
&\leq \tau_{\max}\{\mathscr{I}(\boldsymbol{\theta}^0)\} \lim_{n\to\infty} \sum_{i=1}^{n} \left(\mathrm{E}\left(\|\boldsymbol{L}_i\|^2 \right)^2 \times \mathrm{E}\left(\mathbb{1}_{\|\boldsymbol{L}_i\|^2 > \varepsilon \tau_{\max}\{\mathscr{I}(\boldsymbol{\theta}^0)\}^{-1}} \right)^2 \right)^{1/2} \\
&= \lim_{n\to\infty} O_p\left(\frac{d^{3/2}}{n^{1/2}} \right) = 0
\end{aligned}
$$

Thus the Lindeberg condition is satisfied and we can apply the multivariate Lindeberg-Feller central limit theorem to obtain $n^{-1/2} \boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^0) \xrightarrow{d} \mathscr{N}(\mathbf{0}, \boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top)$, and

$$
n^{1/2} \boldsymbol{G} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathscr{N}(\mathbf{0}, \boldsymbol{G} \mathscr{I}^{-1}(\boldsymbol{\theta}^0) \boldsymbol{G}^\top).
$$

Taking $\boldsymbol{G}$ as the matrix which identifies the elements of $\boldsymbol{\kappa}$ in $\boldsymbol{\theta}$, recalling that $\dim(\boldsymbol{\kappa}) = p$ is finite, it follows that $n^{1/2} \left(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}^0 \right) \xrightarrow{\mathrm{d}} \mathscr{N}(\mathbf{0}, \mathscr{I}^{-1}(\boldsymbol{\theta}^0)_{\boldsymbol{\kappa}})$ as required.

B Illustration of How to Construct P-spline Basis Functions

We provide a simple illustration of how to construct P-spline (penalized B-spline) basis functions and perform the centering constraint. A general overview of spline basis construction as part of GAMs can be found in Sections 4.1 and 4.2 of Wood (2006), while more technical details on P-splines may be found in Eilers and Marx (1996), Poliakoff et al. (1999) and De Boor (2001).

Consider covariate $j = 1, \ldots, q$, for which we want to construct a B-spline basis of degree $m_j = 3$ and $K_j = 5$ interior knots. In what follows, we shall suppress the dependence on j for ease of notation. With K interior knots, we define a total of $K + 2m + 2$ knot locations $u_0 < u_1 < u_2 < \cdots < u_{K+2m+1}$, with the first and last m knots placed at essentially arbitrary locations. Generally, the k^{th} B-spline basis function of degree m, written as $Z_{k,m}(u)$, can be defined recursively as follows. For $k = 0, \ldots, K + 2m + 1$,

$$Z_{k,m}(u) = \frac{u - u_k}{(u_{k+m} - u_k)} Z_{k,m-1}(u) + \frac{u_{k+m+1} - u}{u_{k+m+1} - u_{k+1}} Z_{k+1,m-1}(u),$$

and

$$Z_{k,0}(u) = \begin{cases} 1 & \text{if } u_k \leq u < u_{k+1}; \\ 0 & \text{if otherwise.} \end{cases}$$

The above recursion is referred to as the Cox-de Boor recursion formula (De Boor, 2001). Note that $m = 0$ corresponds to a step function assigning one if u is in the k^{th} knot span $[u_k, u_{k+1})$. When $m = 1$, $Z_{k,1}(u)$ spans two intervals and is piecewise linear while $Z_{k,2}(u)$ spans three intervals and is piecewise quadratic, $Z_{k,3}(u)$ spans four intervals that is piecewise cubic, and so on. As discussed in the main text, the most common choice is cubic B-splines with $m = 3$, as it provides sufficient flexibility for approximating most functions. Note the local support of B-splines is one of main computational advantages, in contrast to standard polynomial splines, for instance.

At value u_i for observation $i = 1, \ldots, n$ the B-spline basis function smooth is then given by

$$s(u_i) = \sum_{k=0}^{K+m} Z_{k,m}(u_i)\beta_k = \boldsymbol{z}_i^\top \boldsymbol{\beta},$$

as given in Section 2 of the main text, where $\boldsymbol{\beta} = (\beta_1^\top, \ldots, \beta_d^\top)^\top$ is the vector of smoothing coefficients of dimension $d = K + m + 1$.

Note that the first basis function $k = 0$ is often referred to as the "intercept", and in typical B-spline implementations is removed to avoid multicollinearity issues. This in turn leaves $\boldsymbol{z}_i$ being of dimension $d = K + m$. In the below, we assume that this has been done.

P-splines are formed by combining the above B-spline basis functions with a difference penalty on the smoothing coefficients to control for overfitting. Typically this is based on squared difference of adjacent coefficients, such that the penalty can be written in the form $P = \lambda \sum_{k=1}^{K+m} (\beta_{k+1} - \beta_k)^2$. This however can also be written in matrix form

$$\begin{aligned} P &= \lambda \boldsymbol{\beta}^\top \begin{pmatrix} 1 & -1 & 0 & \ldots \\ -1 & 2 & -1 & \ldots \\ 0 & -1 & 2 & \ldots \\ \vdots & \vdots & \vdots & \ldots \end{pmatrix} \boldsymbol{\beta} \\ &= \lambda \boldsymbol{\beta}^\top \boldsymbol{S} \boldsymbol{\beta}, \end{aligned}$$

where $\lambda > 0$ is the smoothing parameter.

The construction of P-splines outlined above does not enforce any centering constraint, which is reflected in the fact that both the $n \times (K + m)$ B-spline model matrix $\boldsymbol{Z}$ is only of rank $(K + m - 1)$. To enforce this constraint, we want the mean of the elements $\boldsymbol{Z}\boldsymbol{\beta}$ to be zero, which is equivalent to wanting $\mathbf{1}^\top \boldsymbol{Z}\boldsymbol{\beta} = 0$ where $\mathbf{1}$ is a $n-$vector of ones. This constraint can be achieved via reparameterization: specifically, if we can find a $(K + m) \times (K + m - 1)$ orthogonal matrix $\boldsymbol{H}$ such that $\mathbf{1}^\top \boldsymbol{Z}\boldsymbol{H} = \mathbf{0}$, then we can set $\boldsymbol{\beta} = \boldsymbol{H}\tilde{\boldsymbol{\beta}}$ for a new vector of smoothing coefficients $\tilde{\boldsymbol{\beta}}$ of

length $(K+m-1)$ such that $\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{H}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\beta}}$ automatically satisfies the centering constraint. Note that in turn, the smoothing penalty is redefined as $\boldsymbol{\beta}^\top \boldsymbol{S}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}^\top \boldsymbol{H}^\top \boldsymbol{S}\boldsymbol{H}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{S}}\tilde{\boldsymbol{\beta}}$ and the penalization is on the reparameterized smoothing coefficients $\tilde{\boldsymbol{\beta}}$. Such an orthogonal matrix can be found by applying a QR decomposition to the vector of columns sums of $\boldsymbol{Z}$ and then taking the last $(K+m-1)$ columns of the associated Q matrix.

In the script `pspline_set_up.R`, we present code which demonstrates how the above is implemented in `R`. This code is used as part of our main fitting function `vagam` in the script `vagam_main.R`.

C Additional Simulation Results

1) VA assuming an unstructured covariance for $\boldsymbol{A}$ (VA-Unstruc); 2) VA assuming a block diagonal structure $\boldsymbol{A}$ (VA-Bdiag); 3) A penalized likelihood approach using `mgcv`, with all settings set at the default options (mgcv-Default); 4) `mgcv` using P-splines and all other settings set at the default (mgcv-P–splines); 5) A mixed model approach using `gamm4` with P-splines and all other settings set at the default (gamm4). Note that methods 1, 2, and 5 employ a mixed model framework for GAMs, while methods 3 and 4 employ a penalized likelihood framework.

We used a variety of criteria to assess performance, as discussed in Section 6 in the main text.

C.1 Poisson Responses

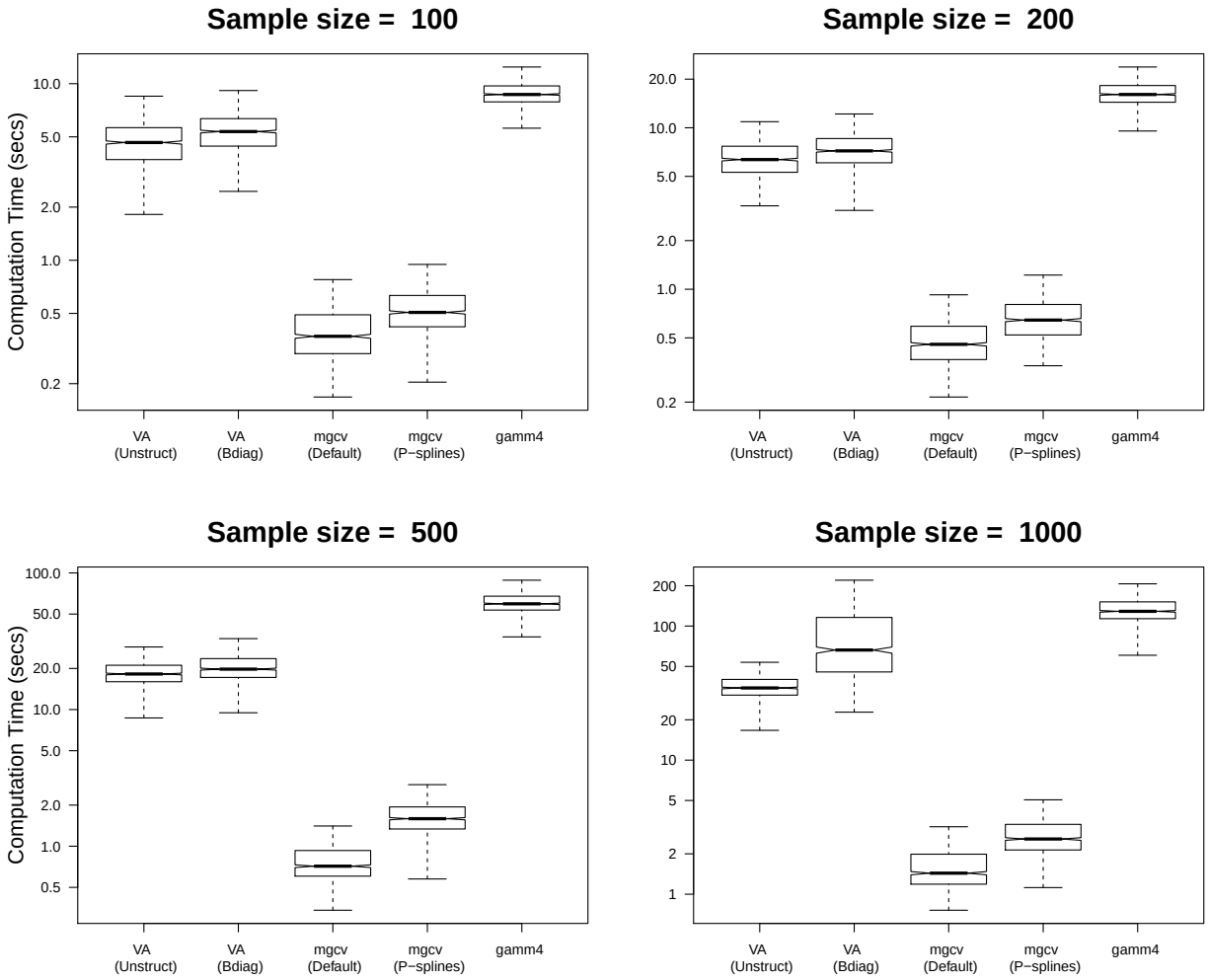
Table 1: Results for Poisson GAMs, based on averaging across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), coverage probability and mean width of the 95% confidence interval for the parametric component ($\text{CI coverage}_\text{p}$ and CI width_p), mean squared error for the overall fit on the linear predictor scale and mean response scale (MSE and MSE_resp), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.007	-0.005	-0.009	-0.010	-0.008
	MSE_p	0.012	0.014	0.010	0.016	0.012
	$\text{CI coverage}_\text{p}$	0.972	0.949	0.936	0.957	0.975
	CI width_p	0.545	0.505	0.348	0.473	0.474
	MSE	0.328	0.330	17.581	1346.472	0.627
	MSE_resp	22.820	22.786	21.021	23.329	22.803
	CI width	2.371	2.092	5.358	90.981	3.350
	Interval score	33.331	33.938	35.469	119.501	32.051
200	Bias_p	-0.004	-0.004	-0.003	-0.005	-0.004
	MSE_p	0.003	0.003	0.002	0.003	0.003
	$\text{CI coverage}_\text{p}$	0.967	0.959	0.941	0.948	0.961
	CI width_p	0.224	0.217	0.167	0.205	0.213
	MSE	0.205	0.203	10.197	36.133	0.395
	MSE_resp	19.619	19.297	16.647	20.103	19.615
	CI width	1.654	1.548	2.511	3.153	2.221
	Interval score	34.887	35.154	35.692	35.190	34.169
500	Bias_p	0.000	0.000	0.000	-0.001	0.000
	MSE_p	0.001	0.001	0.001	0.001	0.001
	$\text{CI coverage}_\text{p}$	0.946	0.943	0.930	0.938	0.941
	CI width_p	0.103	0.101	0.084	0.097	0.100
	MSE	0.125	0.124	2.720	1145.791	0.206
	MSE_resp	13.404	13.198	9.583	13.503	13.396
	CI width	1.179	1.143	1.281	5.359	1.448
	Interval score	36.594	36.696	37.366	40.420	36.258
1000	Bias_p	-0.000	-0.000	-0.000	-0.000	-0.000
	MSE_p	0.000	0.000	0.000	0.000	0.000
	$\text{CI coverage}_\text{p}$	0.968	0.963	0.951	0.965	0.965
	CI width_p	0.061	0.060	0.054	0.059	0.060
	MSE	0.076	0.076	0.430	0.201	0.111
	MSE_resp	9.250	9.110	6.472	9.445	9.244
	CI width	0.883	0.867	0.791	0.949	1.040
	Interval score	37.386	37.398	37.919	37.310	37.202

Table 2: Results for Poisson GAMs, based on taking the median across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), mean width of the 95% confidence interval for the parametric component (CI width_p), mean squared error for the overall fit on the linear predictor scale and mean response scale (MSE and MSE_resp), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.000	-0.005	-0.005	-0.009	-0.002
	MSE_p	0.005	0.005	0.003	0.007	0.005
	CI width_p	0.483	0.444	0.333	0.452	0.465
	MSE	0.285	0.287	0.838	0.402	0.451
	MSE_resp	17.802	17.611	16.217	18.282	17.721
	CI width	2.327	2.035	2.841	2.928	3.134
	Interval score	34.176	34.084	34.395	33.845	32.096
200	Bias_p	-0.003	-0.003	-0.004	-0.006	-0.003
	MSE_p	0.001	0.001	0.001	0.001	0.001
	CI width_p	0.214	0.209	0.165	0.202	0.209
	MSE	0.180	0.179	0.530	0.224	0.267
	MSE_resp	16.461	16.241	13.266	16.997	16.469
	CI width	1.627	1.514	1.743	1.868	2.087
	Interval score	34.163	36.854	36.563	34.109	34.158
500	Bias_p	-0.001	-0.001	0.000	-0.001	-0.001
	MSE_p	0.000	0.000	0.000	0.000	0.000
	CI width_p	0.101	0.100	0.083	0.097	0.099
	MSE	0.101	0.101	0.357	0.118	0.130
	MSE_resp	12.192	11.973	8.546	12.264	12.172
	CI width	1.150	1.116	1.036	1.209	1.381
	Interval score	37.125	37.114	37.204	37.126	37.212
1000	Bias_p	-0.000	-0.001	-0.000	-0.001	-0.000
	MSE_p	0.000	0.000	0.000	0.000	0.000
	CI width_p	0.060	0.060	0.054	0.059	0.060
	MSE	0.061	0.061	0.335	0.066	0.076
	MSE_resp	8.523	8.384	5.664	8.665	8.494
	CI width	0.859	0.843	0.735	0.882	0.996
	Interval score	36.986	36.985	37.420	37.013	37.098

Figure 1: Comparative boxplots of computation time in seconds for various methods of estimating GAMs with Poisson responses. Note time on the y-axis is on the log scale. Outliers have also been removed to better visualize the differences in time between the methods.



C.2 Bernoulli responses

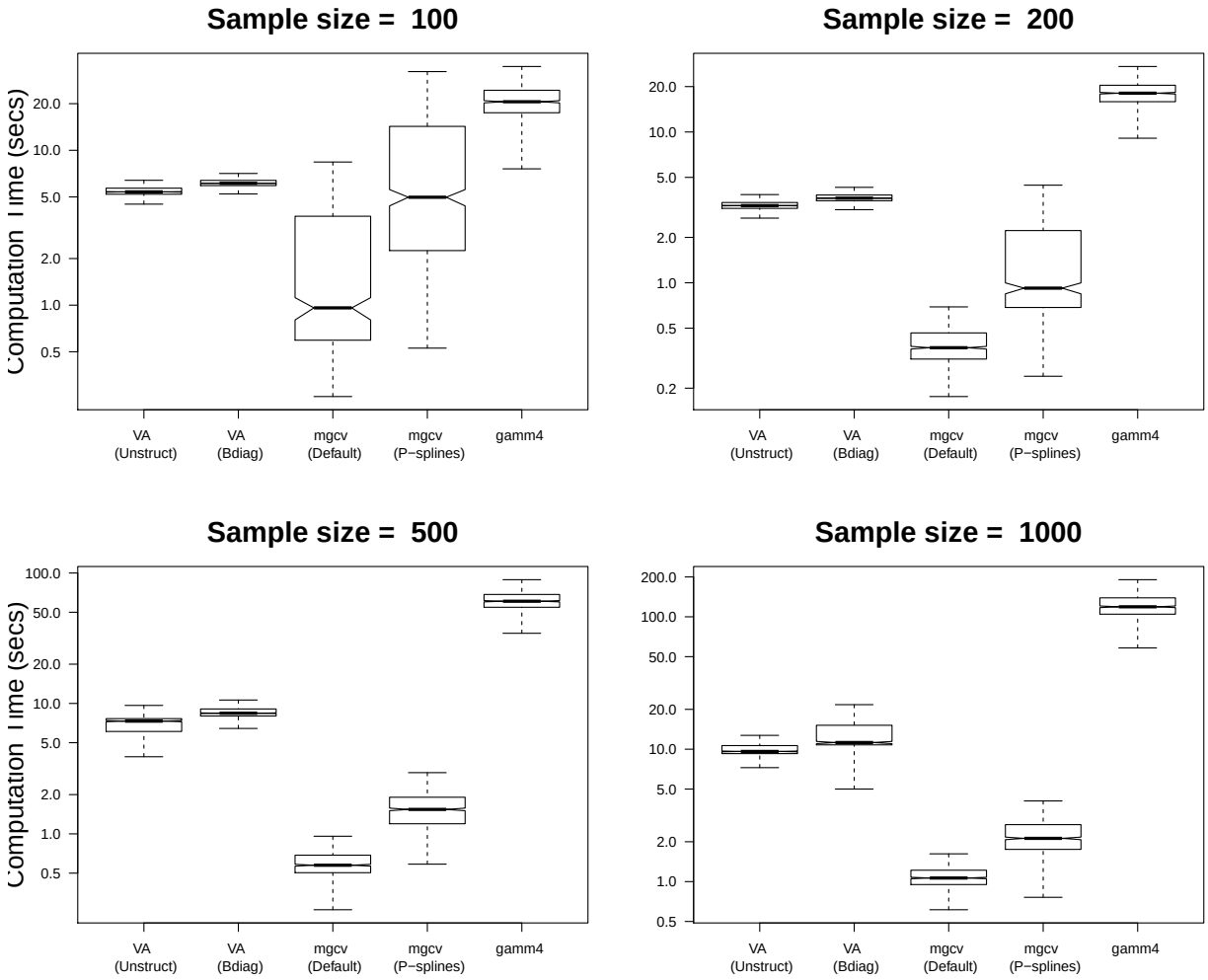
Table 3: Results for Bernoulli GAMs, based on averaging across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), coverage probability and mean width of the 95% confidence interval for the parametric component ($\text{CI coverage}_\text{p}$ and CI width_p), mean squared error for the overall fit on the linear predictor scale and mean response scale (MSE and MSE_resp), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.072	-0.075	10.666	12.417	0.305
	MSE_p	0.319	0.313	2069.048	1468.879	5.934
	$\text{CI coverage}_\text{p}$	0.964	0.965	0.960	0.990	0.934
	CI width_p	2.340	2.327	($> 10^4$)	($> 10^4$)	3.989
	MSE	2.021	2.070	($> 10^4$)	($> 10^4$)	33.948
	MSE_resp	0.020	0.021	0.063	0.085	0.022
	CI width_s	2.690	2.590	($> 10^4$)	($> 10^4$)	10.473
	Interval score	31.466	31.690	($> 10^4$)	($> 10^4$)	30.305
200	Bias_p	-0.060	-0.061	1.471	6.638	0.040
	MSE_p	0.174	0.173	712.480	908.577	0.269
	$\text{CI coverage}_\text{p}$	0.963	0.965	0.928	0.951	0.931
	CI width_p	1.692	1.687	255.685	4166.608	1.862
	MSE	1.191	1.209	($> 10^4$)	($> 10^4$)	1.417
	MSE_resp	0.012	0.012	0.018	0.037	0.012
	CI width_s	2.234	2.168	3894.705	($> 10^4$)	4.567
	Interval score	33.138	33.356	3918.445	($> 10^4$)	29.235
500	Bias_p	-0.032	-0.033	0.019	0.009	0.023
	MSE_p	0.067	0.067	0.083	0.082	0.083
	$\text{CI coverage}_\text{p}$	0.962	0.961	0.956	0.958	0.952
	CI width_p	1.083	1.081	1.153	1.161	1.141
	MSE	0.635	0.663	17.013	($> 10^4$)	0.650
	MSE_resp	0.007	0.007	0.006	0.008	0.007
	CI width_s	1.718	1.669	3.251	8.954	3.166
	Interval score	35.086	35.189	33.735	37.202	32.158
1000	Bias_p	-0.028	-0.029	-0.001	-0.012	0.006
	MSE_p	0.037	0.037	0.041	0.039	0.042
	$\text{CI coverage}_\text{p}$	0.952	0.954	0.961	0.957	0.952
	CI width_p	0.772	0.771	0.798	0.795	0.798
	MSE	0.356	0.379	0.619	0.566	0.350
	MSE_resp	0.004	0.004	0.003	0.004	0.004
	CI width_s	1.368	1.335	2.073	2.445	2.381
	Interval score	35.932	36.079	34.997	33.685	33.765

Table 4: Results for Bernoulli GAMs, based on taking the median across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), mean width of the 95% confidence interval for the parametric component (CI width_p), mean squared error for the overall fit on the linear predictor scale and mean response scale (MSE and MSE_resp), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.071	-0.071	0.669	3.576	0.117
	MSE_p	0.133	0.131	5.915	168.796	0.286
	CI width_p	2.317	2.307	36.566	634.870	2.718
	MSE	1.948	1.992	5267.241	($> 10^4$)	2.004
	MSE_resp	0.019	0.020	0.064	0.093	0.020
	CI width_s	2.627	2.519	2139.575	($> 10^4$)	5.908
	Interval score	31.051	31.148	2141.575	($> 10^4$)	27.103
200	Bias_p	-0.067	-0.067	0.061	0.121	0.031
	MSE_p	0.069	0.068	0.120	0.255	0.099
	CI width_p	1.684	1.681	1.989	2.089	1.836
	MSE	1.127	1.152	3.040	3.185	1.109
	MSE_resp	0.011	0.011	0.015	0.019	0.012
	CI width_s	2.171	2.118	5.889	7.612	4.308
	Interval score	34.020	34.010	31.908	35.013	28.869
500	Bias_p	-0.029	-0.032	0.023	0.016	0.034
	MSE_p	0.028	0.028	0.038	0.035	0.038
	CI width_p	1.080	1.078	1.146	1.146	1.135
	MSE	0.595	0.630	0.679	0.633	0.550
	MSE_resp	0.006	0.007	0.006	0.007	0.006
	CI width_s	1.685	1.624	2.870	3.284	3.079
	Interval score	34.182	36.247	34.448	31.690	31.564
1000	Bias_p	-0.027	-0.027	-0.002	-0.006	0.011
	MSE_p	0.017	0.017	0.017	0.017	0.017
	CI width_p	0.772	0.770	0.796	0.794	0.797
	MSE	0.328	0.356	0.351	0.324	0.311
	MSE_resp	0.004	0.004	0.003	0.004	0.004
	CI width_s	1.337	1.306	2.001	2.340	2.342
	Interval score	37.256	37.240	34.364	34.291	34.320

Figure 2: Comparative boxplots of computation time in seconds for various methods of estimating GAMs with Bernoulli responses. Note time on the y-axis is on the log scale. Outliers have also been removed to better visualize the differences in time between the methods.



C.3 Normal responses

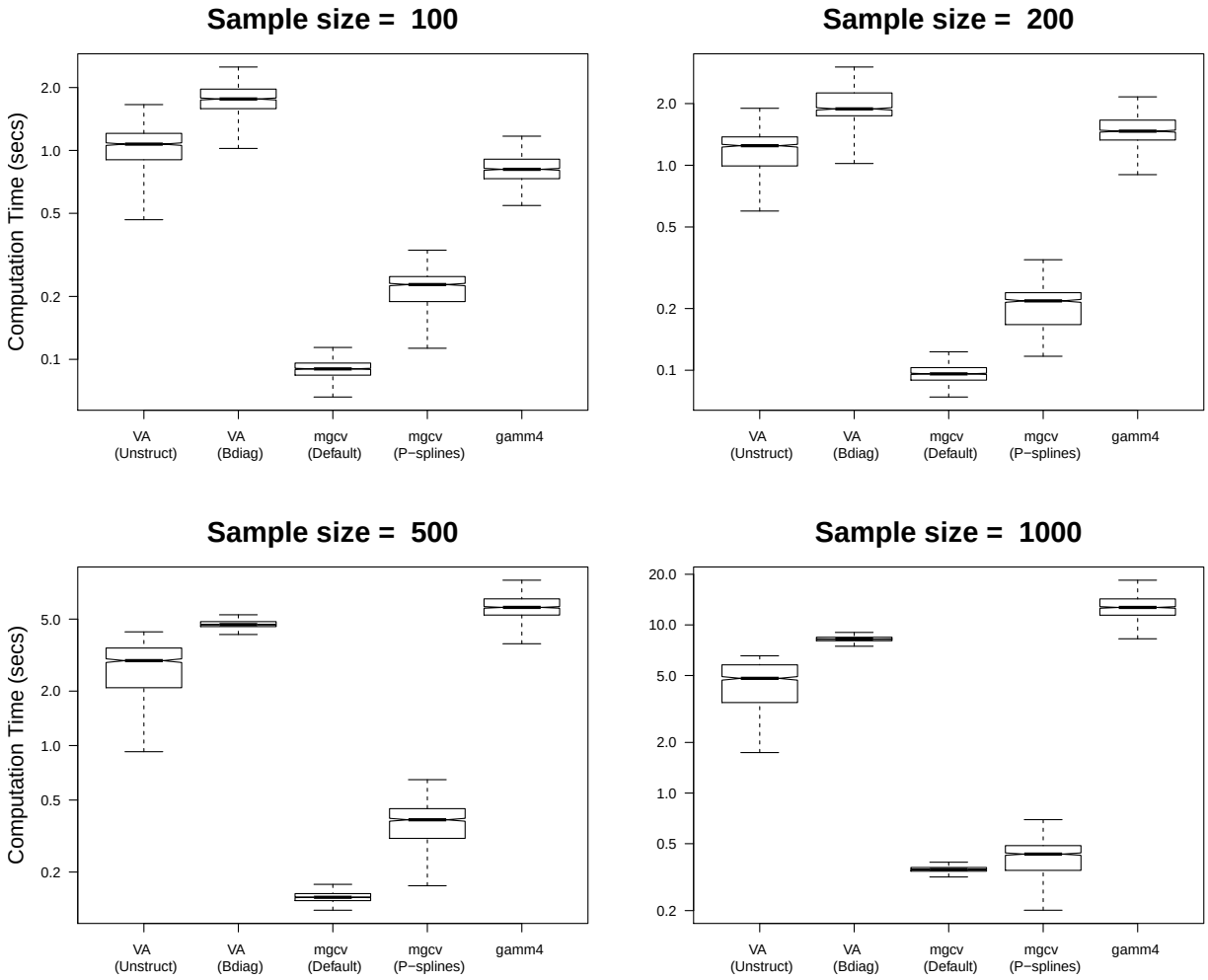
Table 5: Results for normal GAMs, based on averaging across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), coverage probability and mean width of the 95% confidence interval for the parametric component ($\text{CI coverage}_\text{p}$ and CI width_p), mean squared error for the overall fit on the linear predictor scale (MSE; note this is same as the MSE on the mean response scale), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.013	-0.011	-0.010	-0.015	-0.012
	MSE_p	0.188	0.188	0.190	0.197	0.187
	$\text{CI coverage}_\text{p}$	0.932	0.941	0.937	0.933	0.937
	CI width_p	1.658	1.668	1.647	1.646	1.655
	MSE	0.671	0.687	0.709	0.771	0.675
	CI width	3.355	3.094	3.316	3.598	3.534
	Interval score	30.963	31.666	31.488	30.466	30.546
200	Bias_p	0.003	0.002	0.002	0.001	0.003
	MSE_p	0.084	0.084	0.085	0.084	0.084
	$\text{CI coverage}_\text{p}$	0.947	0.948	0.947	0.948	0.946
	CI width_p	1.148	1.154	1.140	1.143	1.147
	MSE	0.389	0.397	0.389	0.416	0.390
	CI width	2.535	2.386	2.331	2.587	2.642
	Interval score	32.695	33.046	33.255	32.547	32.406
500	Bias_p	-0.002	-0.002	-0.003	-0.002	-0.002
	MSE_p	0.035	0.035	0.034	0.035	0.035
	$\text{CI coverage}_\text{p}$	0.950	0.951	0.949	0.951	0.950
	CI width_p	0.715	0.715	0.711	0.713	0.715
	MSE	0.201	0.196	0.180	0.203	0.200
	CI width	1.875	1.802	1.506	1.825	1.925
	Interval score	34.743	34.898	35.790	34.817	34.593
1000	Bias_p	0.004	0.004	0.004	0.004	0.004
	MSE_p	0.015	0.015	0.015	0.015	0.015
	$\text{CI coverage}_\text{p}$	0.956	0.958	0.961	0.956	0.955
	CI width_p	0.501	0.502	0.500	0.501	0.501
	MSE	0.115	0.112	0.103	0.113	0.115
	CI width	1.412	1.368	1.089	1.355	1.443
	Interval score	36.080	36.228	36.937	36.151	35.995

Table 6: Results for normal GAMs, based on taking the median across simulated datasets. Below, we present results for the: bias and mean squared error of the parametric component (Bias_p and MSE_p), mean width of the 95% confidence interval for the parametric component (CI width_p), mean squared error for the overall fit on the linear predictor scale (MSE; note this is same as the MSE on the mean response scale), mean width of the 95% confidence interval and interval score for the ten out of sample validation points (CI width and Interval score).

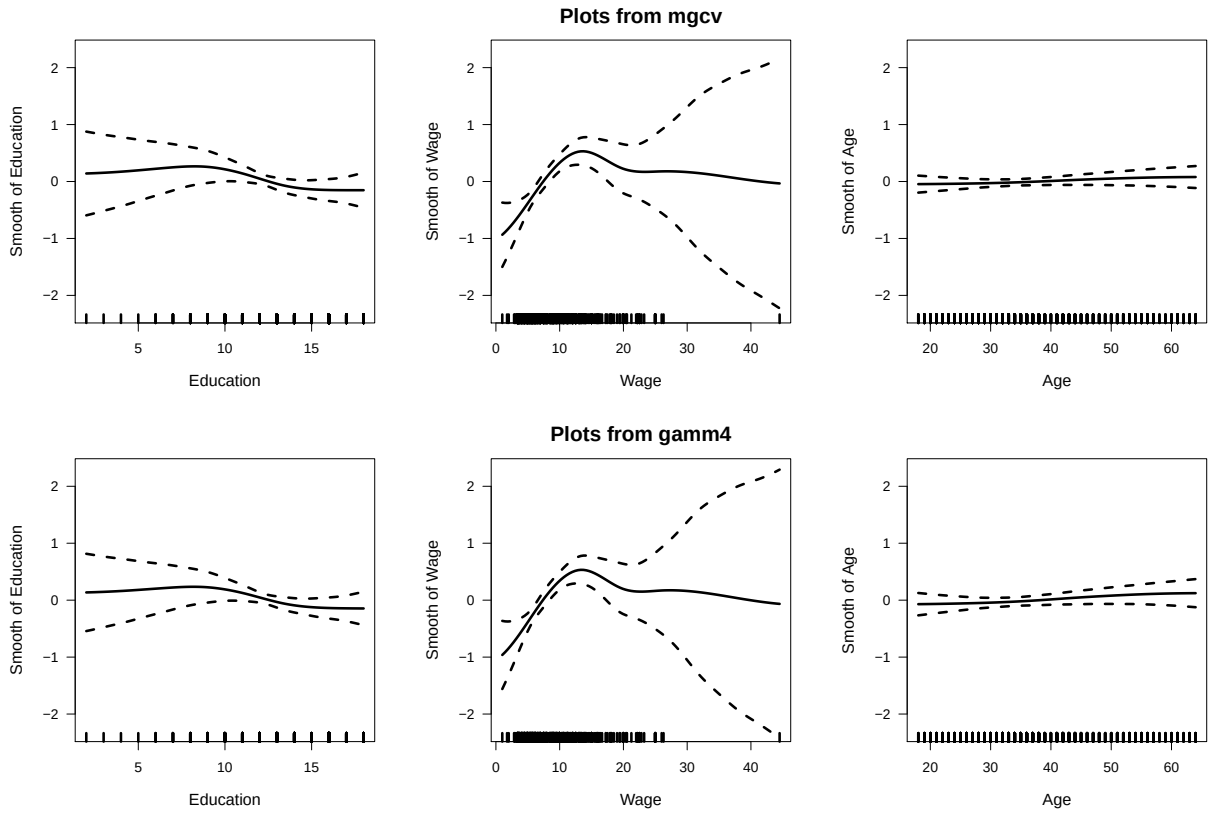
n		VA (Unstruc)	VA (Bdiag)	mgcv (Default)	mgcv (P-splines)	gamm4
100	Bias_p	-0.018	-0.020	-0.003	-0.028	-0.022
	MSE_p	0.082	0.085	0.085	0.081	0.084
	CI width_p	1.650	1.662	1.645	1.641	1.652
	MSE	0.649	0.663	0.661	0.709	0.652
	CI width	3.337	3.065	3.272	3.509	3.522
	Interval score	31.323	31.239	31.355	31.297	31.346
100	Bias_p	0.005	0.008	-0.004	0.001	0.007
	MSE_p	0.039	0.038	0.037	0.039	0.038
	CI width_p	1.144	1.149	1.140	1.142	1.144
	MSE	0.374	0.377	0.370	0.391	0.374
	CI width	2.523	2.379	2.310	2.565	2.631
	Interval score	34.255	34.226	34.190	34.149	31.179
100	Bias_p	-0.010	-0.011	-0.008	-0.009	-0.009
	MSE_p	0.017	0.016	0.017	0.017	0.017
	CI width_p	0.715	0.716	0.711	0.714	0.715
	MSE	0.196	0.190	0.173	0.195	0.195
	CI width	1.870	1.796	1.498	1.808	1.919
	Interval score	33.947	33.890	37.387	33.941	33.979
100	Bias_p	0.003	0.003	0.003	0.004	0.003
	MSE_p	0.007	0.007	0.007	0.007	0.007
	CI width_p	0.501	0.501	0.500	0.501	0.501
	MSE	0.114	0.109	0.100	0.110	0.113
	CI width	1.407	1.362	1.077	1.344	1.436
	Interval score	37.362	37.325	37.092	37.289	37.386

Figure 3: Comparative boxplots of computation time in seconds for various methods of estimating GAMs with normal responses. Note time on the y-axis is on the log scale. Outliers have also been removed to better visualize the differences in time between the methods.



D Additional Results for Application

Figure 4: Smooths from the fitted GAM using `mgcv` with default settings (top) and `gamm4` with default settings (bottom), regressing union membership as a function of six covariates. In both fitted GAMs, results show that no evidence of a relationship between probability of a worker being in a union and their age, borderline evidence of a relationship to education, and a strong non-linear relationship with their hourly wage.



References

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York.

De Boor, C. (2001). Calculation of the smoothing spline with weighted roughness measure. *Mathematical Models and Methods in Applied Sciences*, 11:33–41.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, 11:89–121.

Poliakoff, J. F., Wong, Y. K., and Thomas, P. D. (1999). An automated curve fairing algorithm for cubic *B*-spline ccurve. *Journal of Computational and Applied Mathematics*, 102(1):73–85.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC press, Boca Raton, FL.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.