Electronic Journal of Statistics
Vol. 9 (2015) 1913–1938
ISSN: 1935-7524
DOI: 10.1214/15-EJS1061

Multivariate sharp quadratic bounds via $\mathbf{\Sigma}$-strong convexity and the Fenchel connection

Ryan P. Browne and Paul D. McNicholas

Department of Mathematics and Statistics, McMaster University, Ontario, Canada

Abstract: Sharp majorization is extended to the multivariate case. To achieve this, the notions of σ-strong convexity, monotonicity, and one-sided Lipschitz continuity are extended to $\mathbf{\Sigma}$-strong convexity, monotonicity, and Lipschitz continuity, respectively. The connection between a convex function and its Fenchel-Legendre transform is then developed. Sharp majorization is illustrated in single and multiple dimensions, and we show that these extensions yield improvements on bounds given within the literature. The new methodology introduced herein is used to develop a variational approximation for the Bayesian multinomial regression model.

Received May 2014.

1. Introduction

With the ever increasing importance of computational tasks in statistics, the class of majorization-maximization algorithms is becoming ever more relevant [7, 10, 14, 11]. Suppose a complicated objective function f is to be minimized. This can be achieved iteratively by constructing a majorizing function g at the current solution x^k and finding a new solution x^{k+1} by minimizing the majorization function. Such an algorithm is an MM algorithm of the majorization-minimization variety, cf. [11]. A function g majorizes a function f at a point y if $g(y) = f(y)$ and $g(x) \geq f(x)$ for all $x \neq y$. Majorization-minimization algorithms are useful when a majorizing function g is easier to minimize than the original objective function f.

[7] present a detailed examination of majorization-minimization with univariate quadratic majorizing functions, i.e., they find sharp quadratic univariate majorizers that are ‘closest’ to the original function. In this paper, the connection between majorization of f and minorization of f^*, the Fenchel-Legendre transform of f, is illustrated. Notably, this is done both in the univariate and multivariate cases. Before this connection can be established, the notions of σ-strong convexity, monotonicity, and one-sided Lipschitz continuity are extended to $\mathbf{\Sigma}$-strong convexity, monotonicity, and Lipschitz continuity, respectively.

First, the notions of σ-strong convexity and monotonicity are presented (Section 2.1). Section 2.2 extends these definitions and gives a theorem that establishes the connection amongst the concepts. In Section 3, the relationship between quadratic majorization of f and quadratic minorization of f^* is proved.

The principles in Sections 2.2 and 3 are then illustrated via univariate and multivariate examples (Section 4). In Section 5, the new methodology is used to develop a variational approximation for Bayesian multinomial regression, and the resulting method is used to analyze data on the prevalence of pneumoconiosis among coalminers. Then a simulation is carried out to study the computational efficiency of our approach (Section 6), and the paper concludes with a brief discussion (Section 7).

2. Methodology

2.1. Background definitions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if

$$f((1-\alpha)\mathbf{y} + \alpha\mathbf{x}) \leqslant (1-\alpha)f(\mathbf{y}) + \alpha f(\mathbf{x}) \tag{1}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha \in (0, 1)$. Given a differentiable convex function f and expansion point $\mathbf{y}$, a supporting hyperplane can be constructed such that

$$f(\mathbf{x}) \geqslant f(\mathbf{y}) + \nabla f(\mathbf{x})\,(\mathbf{x} - \mathbf{y}) \tag{2}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

For any function $f : \mathbb{R}^n \to \mathbb{R}$, the Legendre-Fenchel transform or conjugate function is the function $f^* : \mathbb{R}^n \to \mathbb{R}$ given by

$$f^*(\mathbf{p}) := \sup_{\mathbf{x}} \{\mathbf{p}'\mathbf{x} - f(\mathbf{x})\}\,. \tag{3}$$

Note that the function f^* is convex and closed regardless of whether f is convex because it is the intersection of the epigraphs of the linear functions of $\mathbf{p}$. The conjugate of the conjugate, f^{**}, will not be the original function f; however, by the conjugacy theorem, if f is convex and closed then $f^{**} = f$. The Fenchel inequality shows that the functions f and f^* satisfy

$$f(\mathbf{x}) + f^*(\mathbf{p}) \geq \mathbf{p}'\mathbf{x}$$

for all $\mathbf{x}, \mathbf{p}$; cf. [17] for a general discussion of the Legendre-Fenchel transform.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *σ-strongly convex* if there is a constant $\sigma > 0$ such that

$$f((1-\alpha)\mathbf{y} + \alpha\mathbf{x}) \leqslant (1-\alpha)f(\mathbf{y}) + \alpha f(\mathbf{x}) - \frac{\sigma}{2}\alpha(1-\alpha) \parallel \mathbf{y} - \mathbf{x} \parallel^2 \tag{4}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ and $\alpha \in (0, 1)$. A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is called *monotone* if it has the property that

$$(T(\mathbf{x}) - T(\mathbf{y}))'\,(\mathbf{x} - \mathbf{y}) = (\mathbf{p} - \mathbf{q})'\,(\mathbf{x} - \mathbf{y}) \geqslant 0 \tag{5}$$

whenever $\mathbf{q} \in T(\mathbf{y})$ and $\mathbf{p} \in T(\mathbf{x})$, and *strictly monotone* if this inequality is strict when $\mathbf{y} \neq \mathbf{x}$. A function f is *σ-strong monotone* if

$$(T(\mathbf{x}) - T(\mathbf{y}))'\,(\mathbf{x} - \mathbf{y}) = (\mathbf{p} - \mathbf{q})'\,(\mathbf{x} - \mathbf{y}) \geqslant \sigma \parallel \mathbf{y} - \mathbf{x} \parallel^2 \tag{6}$$

whenever $\mathbf{q} \in T(\mathbf{y})$ and $\mathbf{p} \in T(\mathbf{x})$.

2.2. Definitions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathbf{\Sigma}$-*strongly convex* if there is a positive definite matrix $\mathbf{\Sigma} \succ 0$ such that

$$f((1-\alpha)\mathbf{y} + \alpha\mathbf{x}) \leqslant (1-\alpha)f(\mathbf{y}) + \alpha f(\mathbf{x}) - \frac{1}{2}\alpha(1-\alpha)\left(\mathbf{y}-\mathbf{x}\right)' \mathbf{\Sigma}\left(\mathbf{y}-\mathbf{x}\right) \quad (7)$$

for all $\mathbf{y}, \mathbf{x} \in R^n$ and $\alpha \in (0,1)$.

A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is $\mathbf{\Sigma}$-*strong monotone* if there exists a positive definite matrix $\mathbf{\Sigma} \succ 0$ such that $T - \mathbf{\Sigma}$ is monotone that is

$$\left[(\mathbf{p} - \mathbf{\Sigma x}) - (\mathbf{q} - \mathbf{\Sigma y})\right]' (\mathbf{x} - \mathbf{y}) \geqslant 0 \quad (8)$$

whenever $\mathbf{q} \in T(\mathbf{y})$ and $\mathbf{p} \in T(\mathbf{x})$. Equivalently, we have $(\mathbf{p}-\mathbf{q})'(\mathbf{x}-\mathbf{y}) \geqslant (\mathbf{x}-\mathbf{y})' \mathbf{\Sigma} (\mathbf{x}-\mathbf{y})$.

A function $T : \mathbb{R}^n \to \mathbb{R}^n$ is *one-sided Lipschitz continuous* on C if there exists a positive definite matrix $\mathbf{\Sigma} \succ 0$ such that

$$(T(\mathbf{x}) - T(\mathbf{y}))' (\mathbf{x}-\mathbf{y}) \leqslant (\mathbf{x}-\mathbf{y})' \mathbf{\Sigma} (\mathbf{x}-\mathbf{y}) \quad (9)$$

for all $\mathbf{x}, \mathbf{y} \in C$. In this case, $\mathbf{\Sigma}$ is called the Lipschitz matrix on C. This is an extension of one-sided Lipschitz continuity, cf. [9].

If a function is one-sided Lipschitz continuous with Lipschitz matrix $\mathbf{\Sigma}$ then it is Lipschitz continuous with a constant $\lambda_{\max}(\mathbf{\Sigma})$ because $\lambda_{\max}(\mathbf{\Sigma})\mathbf{I}_p \succ \mathbf{\Sigma}$, where $\lambda_{\max}(\mathbf{\Sigma})$ is the largest eigenvalue of $\mathbf{\Sigma}$. If a mapping is $\mathbf{\Sigma}$-strong monotone then it is σ-strong monotone because $\mathbf{\Sigma} \succ \mathbf{I}_p\lambda_{\min}(\mathbf{\Sigma})$, where $\lambda_{\min}(\mathbf{\Sigma})$ is the smallest eigenvalue of $\mathbf{\Sigma}$. If a mapping is $\mathbf{\Sigma}$-strong convex then it is σ-strong convex because $\mathbf{\Sigma} \succ \mathbf{I}_p\lambda_{\min}(\mathbf{\Sigma})$.

2.3. Theorems

Proofs for the following theorems are given in Appendix A.

Theorem 2.1 ($\mathbf{\Sigma}$-strong Equivalenence)**.** *Given a function f and a positive definite matrix $\mathbf{\Sigma} \succ 0$, the following are equivalent:*

1. *$f(\mathbf{y})$ is $\mathbf{\Sigma}$-strongly convex.*
2. *$h(\mathbf{y}) = f(\mathbf{y}) - \frac{1}{2}\mathbf{y}'\mathbf{\Sigma}\mathbf{y}$ is convex.*
3. *∇f is $\mathbf{\Sigma}$-strongly monotone.*
4. *$f(\mathbf{x}) \geqslant f(\mathbf{y}) + \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})'\mathbf{\Sigma}(\mathbf{x}-\mathbf{y})$, for all $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$.*

Theorem 2.2 ($\mathbf{\Sigma}$-Bounds)**.** *Given a function f and a positive definite matrix $\mathbf{\Sigma} \succ 0$, the following are equivalent:*

1. *∇f is one-sided $\mathbf{\Sigma}$-Lipschitz continuous.*
2. *$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{x})'(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})'\mathbf{\Sigma}(\mathbf{x}-\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*
3. *For some $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}$ and $t \in [0,1]$, f satisfies*

$$f(t\mathbf{x}+(1-t)\mathbf{y}) \geqslant tf(\mathbf{x})+(1-t)f(\mathbf{y}) - \frac{1}{2}t(1-t)\left(\mathbf{x}-\mathbf{y}\right)'\mathbf{\Sigma}\left(\mathbf{x}-\mathbf{y}\right). \quad (10)$$

Theorem 2.3 ($\mathbf{\Sigma}$-strong Dualization)**.** *Given a function f and a positive definite matrix $\mathbf{\Sigma} \succ 0$, the following are equivalent:*

1. *f has a quadratic upper bound*

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})' (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})' \mathbf{\Sigma} (\mathbf{x} - \mathbf{y}), \tag{11}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

2. *f^* has a quadratic lower bound*

$$f(\mathbf{p}) \geqslant f(\mathbf{q}) + \nabla f(\mathbf{q})' (\mathbf{p} - \mathbf{q}) + \frac{1}{2} (\mathbf{p} - \mathbf{q})' \mathbf{\Sigma}^{-1} (\mathbf{p} - \mathbf{q}), \tag{12}$$

for all $\mathbf{q}, \mathbf{p}$.

Corollary 2.4 (The $\mathbf{\Sigma}^{-1}$-strong gradient inequality)**.** *If f^* is $\mathbf{\Sigma}^{-1}$-strong convex, then ∇f and ∇f^* satisfy the inequality*

$$0 \leqslant (\mathbf{p} - \mathbf{q})' \mathbf{\Sigma}^{-1} (\mathbf{p} - \mathbf{q}) \leqslant (\mathbf{p} - \mathbf{q})' (\mathbf{x} - \mathbf{y}) \leqslant (\mathbf{x} - \mathbf{y})' \mathbf{\Sigma} (\mathbf{x} - \mathbf{y}) \tag{13}$$

for any $\mathbf{q} \in \nabla f(\mathbf{y})$ and $\mathbf{p} \in \nabla f(\mathbf{x})$ or, equivalently, $\mathbf{y} \in \nabla f^*(\mathbf{q})$ and $\mathbf{x} \in \nabla f^*(\mathbf{p})$.

Proof. f^* being $\mathbf{\Sigma}^{-1}$-strong convex is equivalent to ∇f^* being $\mathbf{\Sigma}^{-1}$-strong monotone, which is the left side of inequality. For the right side, by combing the $\mathbf{\Sigma}^{-1}$-strong dualization theorem and the $\mathbf{\Sigma}$ bound theorem, this is equivalent to f being one-sided Lipschitiz continuous with a matrix $\mathbf{\Sigma}$. □

3. Sharp quadratic majorization

A quadratic function $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma})$ is a $\mathbf{\Sigma}$-*quadratic majorizor* of $f(\mathbf{x})$ at the point $\mathbf{y}$ if

$$f(\mathbf{x}) \leqslant q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}) = f(\mathbf{y}) + \nabla f(\mathbf{y})' (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})' \mathbf{\Sigma} (\mathbf{x} - \mathbf{y}) \tag{14}$$

for all $\mathbf{x}$ and any $\mathbf{\Sigma} \succ 0$. A quadratic majorizor $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}_*)$ of $f(\mathbf{x})$ at $\mathbf{y}$ is the *sharp quadratic majorizer* if $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}_*) \leqslant q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma})$ for all $\mathbf{x}$ or, equivalently, $\mathbf{\Sigma} \succeq \mathbf{\Sigma}_* \succ 0$.

A quadratic function $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma})$ is a $\mathbf{\Sigma}$-*quadratic minorizor* of $f(\mathbf{x})$ at the point $\mathbf{y}$ if

$$f(\mathbf{x}) \geqslant q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}) = f(\mathbf{y}) + \nabla f(\mathbf{y})' (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})' \mathbf{\Sigma} (\mathbf{x} - \mathbf{y}) \tag{15}$$

for all $\mathbf{x}$ and any $\mathbf{\Sigma} \succ 0$. A quadratic minorizor $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}_*)$ of $f(\mathbf{x})$ at $\mathbf{y}$ is the *sharp quadratic minorizor* if $q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma}_*) \geqslant q(\mathbf{x}|\mathbf{y}, f, \mathbf{\Sigma})$ for all $\mathbf{x}$ or, equivalently, $\mathbf{\Sigma} \succeq \mathbf{\Sigma}_* \succ 0$.

Lemma 3.1 ($\mathbf{\Sigma}$-Sharp Quadratic Dualization (Majorization-Minorization))**.** *A matrix $\mathbf{\Sigma}$ is the sharp quadratic majorizor of a convex function f if and only if $\mathbf{\Sigma}^{-1}$ is the sharp quadratic minorizor of f^*.*

Proof. By contradiction. Assume that $\boldsymbol{\Sigma}$ is the sharp quadratic majorizor of f and $\boldsymbol{\Sigma}_*^{-1}$ is the sharp quadratic minorizor of f^*, where $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_*$. By the strong dualization theorem, $\boldsymbol{\Sigma}^{-1}$ is a quadratic minorizor of f^* and $\boldsymbol{\Sigma}_*$ is a quadratic majorizor of f. Because $\boldsymbol{\Sigma}$ is the sharp quadratic majorizor of f and $\boldsymbol{\Sigma}_*^{-1}$ is the sharp quadratic minorizor of f^*, we have that $\boldsymbol{\Sigma}_* \succeq \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_*^{-1} \succeq \boldsymbol{\Sigma}^{-1}$. However, by the properties of Löwner ordering, if $\boldsymbol{\Sigma}_* \succeq \boldsymbol{\Sigma}$ then $\boldsymbol{\Sigma}^{-1} \succeq \boldsymbol{\Sigma}_*^{-1}$; however, we require $\boldsymbol{\Sigma}_*^{-1} \succeq \boldsymbol{\Sigma}^{-1}$ because $\boldsymbol{\Sigma}_*^{-1}$ is the sharp quadratic minorizor of f^*. This can only be true if $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_*^{-1}$, which is a contradiction. □

4. Examples of sharp quadratic majorization

In this section, we illustrate the above methodology in univariate and multivariate cases. The univariate case illustrates duality with other results within the literature. The multivariate case presents a solution that could not be obtained without the methodology presented herein.

4.1. Univariate example: Univariate logistic function

The Fenchel connection is illustrated and compared with results obtained from [7], who examined sharp quadratic majorization in the univariate case. We begin this comparison with a Taylor expansion with the remainder in integral form for a convex function $f(x)$ at the expansion point y, i.e.,

$$f(x) = f(y) + f'(y)(x-y) + \frac{(x-y)^2}{2} a(x,y),$$

where

$$a(x,y) = 2\int_0^1 (1-t) f''(t)(y + t(x-y))dt.$$

If $a(x,y)$ is bounded, we can form a quadratic majorization function. In sharp quadratic majorization, we are interested in $\sigma(y) = \sup_x a(x,y)$ for a particular y because we can form the sharp majorization function with the property

$$f(x) \leqslant f(y) + f'(y)(x-y) + \frac{(x-y)^2}{2}\,\sigma(y).$$

We are interested in the latter term. If the $\sup_x a(x,y)$ is attained at $z \neq y$, then the results of [7] tell us we have

$$\sigma(y) = \frac{f'(z) - f'(y)}{z-y} \tag{16}$$

and we can see duality via

$$\frac{1}{\sigma} = \frac{z-y}{f'(z)-f'(y)} = \frac{g'(w)-g'(q)}{w-q}, \tag{17}$$

where $g = f^*$, the Fenchel transform of f.

For the univariate logistic function, we have

$$f(x) = \log\left(1+e^{x}\right), \quad f'(x) = \frac{1}{\left(1+e^{-x}\right)}, \quad \text{and} \quad f''(x) = \frac{e^{x}}{\left(1+e^{x}\right)^{2}}. \tag{18}$$

Note, the second derivative is bounded above by 1/4.

[5] suggest using a bound on the second derivative for the binomial function. We have

$$a(x,y) = 2\int_0^1 (1-t)f''(t)(y+t(x-y))dt \le 2\int_0^1 (1-t)\frac{1}{4}dt = \frac{1}{4}.$$

The quadratic majorization function based on the Böhning bound is

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2}\frac{1}{4}(x-y)^2.$$

[12] use convexity as an argument for their bound. [7] show that it can be a one-dimensional sharp quadratic bound. Here, we motive it through minimizing the Taylor remainder for an expansion point y. Using integration by parts, we obtain

$$a(x,y) = \frac{f'(y)(x-y) + f(x) - f(y)}{\frac{1}{2}(x-y)^2}.$$

Taking the derivative and solving gives

$$\sigma(y) = a(-y,y) = \frac{f'(y)(-2y) - y}{2y^2} = \frac{1}{y}\left(\frac{1}{1+e^{-y}} - \frac{1}{2}\right),$$

which is the bound used by [12]. This bound minimizes the remainder term for a quadratic Taylor expansion at a particular expansion point a; thus, it is the sharp quadratic majorization bound.

The Fenchel transform for the univariate logistic function is

$$g(p) = f^{*}(p) = p\log(p) + (1-p)\log(1-p),$$

for $p \in (0,1)$; accordingly,

$$g'(p) = \log(p) - \log(1-p) \qquad \text{and} \qquad g''(p) = \frac{1}{p} + \frac{1}{1-p}.$$

The second derivative is bounded below by 4. The Taylor expansion in remainder form of the Fenchel transform at a given point p is

$$g(p) = g(q) + g'(q)(p-q) + \frac{(p-q)^2}{2}\, b(x,y),$$

where

$$b(p,q) = 2\int_0^1 (1-t)g''(t)(q+t(p-q))dt = \frac{g'(q)(p-q) + g(p) - g(q)}{\frac{1}{2}(p-q)^2}.$$

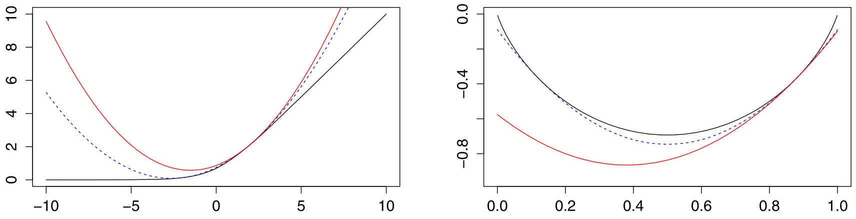

FIG 1. *Plots comparing the bounds on the univariate function* $f(x) = \log(1 + e^x)$ *and its conjugate function* $p\log(p) + (1-p)\log(1-p)$ *using the expansion point* $y = 2$ *or* $q = 0.88$*. The black, dashed-blue, and red lines represent the function, the sharp quadratic bound, and the Böhning quadratic bound, respectively.*

Now, we want the sharp quadratic minorization bound,

$$g(p) \geq g(q) + g'(q)(p-q) + \frac{(p-q)^2}{2}\,\lambda(q),$$

where $\lambda(q) \geq b(p, q)$ for all p. We find that

$$\lambda(q) = b(1-q, q) = \frac{g'(1-q)(1-q-q) + g(1-p) - g(q)}{\frac{1}{2}(1-q-q)^2} = \frac{\log(1-q) - \log(q)}{\frac{1}{2}(1-2q)}.$$

Note that when $q = f'(y)$ or $y = g'(q)$, we have the relation $1/\lambda(q) = \sigma(y)$.

Now, we illustrate and compare the univariate bounds in Figure 1, which is obtained using the expansion point $y = 2$ and the conjugate expansion point $q = e^2/(1+e^2) \approx 0.88$, and then computing the corresponding sharp bounds using the functions $\sigma(y) = 0.19$ and $\lambda(q) = 5.25$. The sharp upper quadratic bound $(y = 2, q = 0.88)$ is

$$\log\left(1 + e^x\right) \leq 2.13 + 0.88(x-2) + \frac{0.19}{2}\left(x-2\right)^2,$$

and the corresponding sharp lower bound for conjugate function using the expansion $q = e^2/(1+e^2) = 0.88$ is

$$p\log\left(p\right) + (1-p)\log\left(1-p\right) \geq -0.37 + 2(x-2) + \frac{5.25}{2}\left(x - 0.88\right)^2.$$

4.2. Multivariate example: Multivariate log-exp function

Here, we illustrate the Fenchel connection in the multivariate setting and then obtain a new quadratic majorization function that cannot be obtained without the methodology introduced herein. Then, this new quadratic majorization function is compared to the quadratic majorization function given by [5].

The multivariate log-exp function is

$$f(\mathbf{x}) = \log\left(1 + \sum_{j=1}^{k} e^{x_j}\right), \tag{19}$$

for $\mathbf{x} \in \mathbb{R}^k$. Its gradient and Hessian are given by

$$\begin{aligned} \nabla f(\mathbf{x}) &= \left(\frac{e^{x_1}}{1+\sum_{j=1}^k e^{x_j}}, \ldots, \frac{e^{x_p}}{1+\sum_{j=1}^k e^{x_j}} \right) \qquad \text{and} \\ \nabla^2 f(\mathbf{x}) &= \operatorname{diag}(\nabla f(\mathbf{x})) - \nabla f(\mathbf{x}) \nabla f(\mathbf{x})', \end{aligned} \tag{20}$$

respectively, and we note that the Hessian is bounded. The Fenchel transform for f is given by

$$f^*(\mathbf{p}) = \sum_{j=1}^k p_j \log p_j + \left(1 - \sum_{j=1}^k p_j\right) \log \left(1 - \sum_{j=1}^k p_j\right), \tag{21}$$

for all $p_j > 0$, $j = 1, \ldots, k$, and $\sum_{j=1}^k p_j < 1$. The gradient and Hessian of the Fenchel transform are given by

$$\begin{aligned} \nabla f^*(\mathbf{p}) &= (\log p_1, \ldots, \log p_k) - \log \left(1 - \sum_{j=1}^k p_j\right) \mathbf{1}_k, \\ &= \left(\log p_1 - \log \left(1 - \sum_{j=1}^k p_j\right), \ldots, \log p_k - \log \left(1 - \sum_{j=1}^k p_j\right) \right), \end{aligned}$$

and

$$\nabla^2 f^*(\mathbf{p}) = \operatorname{diag}\left(\frac{1}{p_1}, \ldots, \frac{1}{p_k}\right) + \left(\frac{1}{1 - \sum_{j=1}^k p_j}\right) \mathbf{J}, \tag{22}$$

respectively, where $\mathbf{J}$ is a $k \times k$ matrix of ones. Note the duality between the Hessians of f and f^*, i.e., $\nabla^2 f(\mathbf{x}) = \left[\nabla^2 f^*(\mathbf{p})\right]^{-1}$, where $\mathbf{p} = \nabla f(\mathbf{x})$ or $\mathbf{x} = \nabla f^*(\mathbf{p})$.

[5] show that the matrix

$$\mathbf{B} = \frac{1}{2}\left[\mathbf{I}_n - \frac{1}{n+1}\mathbf{J}_n\right]$$

has the property that $\mathbf{B} \succeq \nabla^2 f$ and can be used as a quadratic majorization bound in this case.

A multivariate Taylor expansion for a function $f(\mathbf{x})$ at the point $\mathbf{y}$ is

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})' \nabla^2 f(\mathbf{c})(\mathbf{x} - \mathbf{y}), \tag{23}$$

where $\nabla^2 f(\mathbf{c})$ is the Hessian evaluated at some point $\mathbf{c}$ on the line connecting $\mathbf{a}$ to $\mathbf{y}$. A multivariate Taylor expansion with the remainder in integral form is given by

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})' \mathbf{A}(\mathbf{x}, \mathbf{y})(\mathbf{x} - \mathbf{y}),$$

where

$$\mathbf{A}(\mathbf{x}, \mathbf{y}) = 2 \int_0^1 (1-t) \nabla^2 f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right) dt. \tag{24}$$

Note, the integral is with respect to each element of $\mathbf{A}(\mathbf{x}, \mathbf{y})$.

We cannot derive an optimal bound via the direct method here because the integral involving $\nabla^2 f$ in (24) does not have a closed form. However, the integral involving $\nabla^2 f^*$ has a closed form.

4.2.1. Deriving the Böhning bound via the Fenchel connection

The Hessian of the Fenchel transform of f is given in (22). For any vector $\mathbf{v}$, we have

$$\mathbf{v}' \left[\nabla^2 f^*(\mathbf{q})\right] \mathbf{v} = \sum_{j=1}^{n} \frac{v_j^2}{q_j} + \frac{\left(\sum_{j=1}^n v_j\right)^2}{1 - \sum_{j=1}^n q_j}. \tag{25}$$

Now, we put a bound on this expression for any $\mathbf{q} \in S_p$, the p-simplex. One particular bound can be obtained by using a matrix of the same form, namely $\mathbf{B}^{-1} = \mathbf{D} + \lambda \mathbf{J}$ (because we are actually interested in $\mathbf{B}$), where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$. This gives

$$\mathbf{v}'\mathbf{B}^{-1}\mathbf{v} = \mathbf{v}' \left[\mathbf{D} + \lambda \mathbf{J}\right] \mathbf{v} = \sum_{j=1}^{n} d_j v_j^2 + \lambda \left(\sum_{j=1}^{n} v_j\right)^2. \tag{26}$$

We can ensure that $\nabla^2 f^*(\mathbf{q}) \succ \mathbf{B}^{-1}$, if all $d_1 = \cdots = d_n = \lambda = 2$. Therefore, because $\mathbf{B}^{-1}$ is similar to a rank-one update, we can obtain its inverse

$$\mathbf{B}^{-1} = 2\left[\mathbf{I}_n + \mathbf{J}_n\right] \quad \Rightarrow \quad \mathbf{B} = \frac{1}{2}\left[\mathbf{I}_n - \frac{1}{n+1}\mathbf{J}_n\right], \tag{27}$$

which is the Böhning bound derived by [5].

4.2.2. Multivariate Taylor expansion for conjugate function

If $\mathbf{q} = (q_1, \ldots, q_n)$ and $\mathbf{p} = (p_1, \ldots, p_n)$, then

$$\begin{aligned} \mathbf{A}\left(\mathbf{p}, \mathbf{q}\right) &= \left[2 \int_0^1 (1-t) \nabla^2 f^*\left(\mathbf{q} + t(\mathbf{p} - \mathbf{q})\right) dt\right] \\ &= \mathrm{diag}\left(\delta(p_1, q_1), \ldots, \delta(p_n, q_n)\right) + \delta(p_{n+1}, q_{n+1})\mathbf{J}, \end{aligned} \tag{28}$$

where

$$\delta(p, q) = 2\frac{q - p + p\left[\log(p) - \log(q)\right]}{(p-q)^2}, \tag{29}$$

for $j = 1, \ldots, n$, $p_{n+1} = 1 - \sum_{j=1}^n p_j$, and $q_{n+1} = 1 - \sum_{j=1}^n q_j$. Now, we need to find a matrix $\mathbf{E}(\mathbf{q})$ such that $\mathbf{A}(\mathbf{p}, \mathbf{q}) \succeq \mathbf{E}(\mathbf{q})$ for all $\mathbf{p}$. If we minimize

$\mathbf{A}\left(\mathbf{p},\mathbf{q}\right)$ or minimize each $\delta(p,q)$ with respect to $(p_1,\dots,p_n)$ and $(q_1,\dots,q_n)$, then we obtain the Böhning bound. Another approach is to find such an $\mathbf{E}(\mathbf{p})$ by minimizing each $\delta(p_j,q_j)$ with respect to p_j, for $j=1,\dots,n,n+1$, individually. The function $\delta(p,q)$ is monotone decreasing over (0,1) with respect to p and so a natural bound would be

$$a(q)=\lim_{p\to 1}\delta(p,q)=2\frac{q-1-\log(q)}{\left(1-q\right)^2}. \tag{30}$$

Thus, we can construct $\mathbf{E}(\mathbf{q})=\operatorname{diag}\left(a(q_1),\dots,a(q_n)\right)+a(q_{n+1})\mathbf{J}$. However, if any q_j is bigger than ≈ 0.316, then $a(q)$ will be less than 2 and we know from the derivation of the Böhning bound that a global bound for $\delta(p,q)$ is 2. To adjust our bound we set

$$m(q)=2\times\max\left\{\frac{q-1-\log(q)}{\left(1-q\right)^2},1\right\} \tag{31}$$

and construct $\mathbf{M}(\mathbf{q})$ in a similar manner to $\mathbf{E}(\mathbf{q})$ but with elements given by (31). Note that as the dimensions increase, any particular q_j will be less than ≈ 0.316 implying that $m(q_j)$ will be equal to $a(q_j)$. Note that if $n=1$ then $\mathbf{M}(\mathbf{q})$ is not equal to the bound given by [12].

4.2.3. Comparison to the Böhning bound

The bounds are compared in two dimensions so that they can be plotted as ellipsoids, presenting a nice visual illustration (Figure 2). To generate these plots, we choose points in the conjugate space $\mathbf{p}$ and then generate the quadratic term $\mathbf{A}(\mathbf{p},\mathbf{q})$ for each $\mathbf{p}$ from the expansion point $\mathbf{q}$. Because we are interested in the original space, $\mathbf{A}(\mathbf{p},\mathbf{q})^{-1}$ is plotted. In Figure 2: the matrix $\mathbf{E}(\mathbf{q})^{-1}$, which bounds each $\delta(p_j,q_j)$, is in black; $\mathbf{M}(\mathbf{q})^{-1}$, which adjusts the bounds accordingly, is in blue; and both the multivariate $\mathbf{B}$ and 'univariate' $(1/2)\mathbf{I}$ Böhning bounds are shown in red. From these plots, we see that bounds we derive via the Fenchel connection are at least as tight as the Böhning bounds and, in some cases, are much sharper.

5. A variational approximation for Bayesian multinomial regression

5.1. The variational approximation

In this section, we utilize the methodology from the previous section to develop a variational approximation for Bayesian multinomial regression. This extends the work of [12], who used a variational approximation for Bayesian (binary) logistic regression. This provides an alternative to the frequentist paradigm for this model.

Variational Bayes approximations are an iterative Bayesian alternative to the EM algorithm. They have been used for parameter estimation in a variety

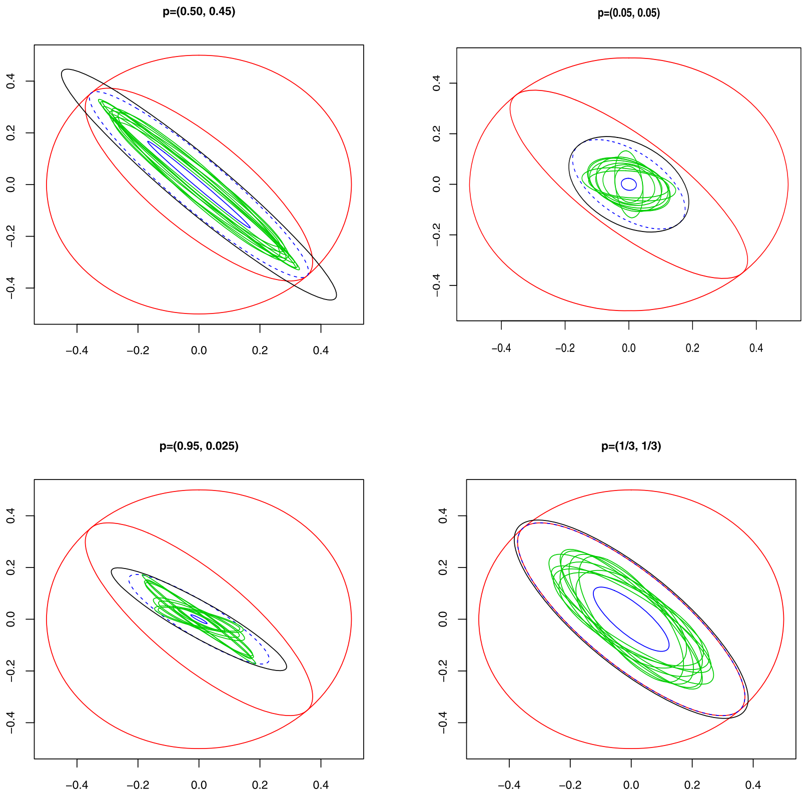

FIG 2. *Plots to compare bounds derived via the Fenchel connection with Böhning bounds. Green ellipsoids represent a sample of* $\mathbf{A}(\mathbf{p},\mathbf{q})^{-1}$ *from randomly selecting 10* $\mathbf{q}$ *uniformly over the 2-simplex. Inner red ellipsoids are Böhning bounds* $\mathbf{B}$*, and outer red ellipsoids are* $\frac{1}{2}\mathbf{I}$*. The matrix* $\mathbf{M}(\mathbf{q})^{-1}$ *is a dashed blue line and the matrix* $\mathbf{E}(\mathbf{q})^{-1}$ *is a black line.*

of settings, including graphical models [13], mixture modelling [6, 19, 16, 18], and mixtures of experts [21]. Variational Bayes approximations have become increasingly popular over the past decade or so due to their fast and deterministic nature as well as the ability to perform simultaneous model selection and parameter estimation. This latter feature circumvents the need for a model selection criterion, which can significantly reduce the associated computational overhead. The joint conditional distribution of the parameters and the missing data is approximated by constructing a tight lower bound on the data marginal likelihood using a computationally convenient density. This approximating density is obtained by minimizing the Kullback–Leibler (KL) divergence between the true and approximating densities [c.f., 3, 16]. Due to the non-negative property of the KL divergence, minimizing the KL divergence is equivalent to maximizing the (tight) lower bound.

We assume a Gaussian prior for the regression coefficients in the Bayesian multinomial regression model. The likelihood function for the multinomial regression model can be constructed by writing a single categorical random variable y, with $k+1$ distinct categories, as a vector $\mathbf{y} = (y_1, \ldots, y_k)$ with dimension k where $y_s = 1$ if y is equal to category s and zero otherwise for $s = 1, \ldots, k$. Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be $n \times k$ matrix of responses and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a $n \times p$ matrix of covariates, then the likelihood for the multinomial regression model can be written

$$g(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}) = \exp\left[\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\Theta}\mathbf{y}_i - \sum_{i=1}^{n} f\left(\mathbf{x}_i'\boldsymbol{\Theta}\right)\right], \tag{32}$$

where $\boldsymbol{\Theta}$ is a $p \times k$ matrix of regression coefficients and the function $g\left(\mathbf{z}\right)$ is the multivariate log-exp sum function. The posterior density of $\boldsymbol{\Theta}$ is

$$g(\boldsymbol{\Theta}|\mathbf{Y}) = \frac{g(\mathbf{Y}, \boldsymbol{\Theta})}{\int g(\mathbf{Y}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}}, \tag{33}$$

which cannot be obtained in closed form. Luckily, the multivariate log-exp sum function is $\boldsymbol{\Sigma}$-strong convex; accordingly, using the methodology developed in the previous section, we can create a surrogate function q based on the expansion point $\boldsymbol{\xi}$, i.e.,

$$\begin{aligned} f(\mathbf{z}) &= \log\left(1 + \sum_{i=1}^{k} e^{z_i}\right) \\ &\leq q(\mathbf{z}|\boldsymbol{\xi}) = f(\boldsymbol{\xi}) - \nabla f(\boldsymbol{\xi})'(\mathbf{z} - \boldsymbol{\xi}) - \frac{1}{2}(\mathbf{z} - \boldsymbol{\xi})'\mathbf{M}(\boldsymbol{\xi})^{-1}(\mathbf{z} - \boldsymbol{\xi}), \end{aligned} \tag{34}$$

where the matrix $\mathbf{M}(\boldsymbol{\xi})$ has the form given in (28) and has elements equal to (31) along with $\mathbf{q} = \nabla f(\boldsymbol{\xi})$ given in (20). The lower bound of the joint distribution $g(\mathbf{Y}, \boldsymbol{\Theta})$ is given by

$$\begin{aligned} \underline{g}(\mathbf{Y}, \mathbf{X}, \boldsymbol{\Theta}|\boldsymbol{\xi}) = \exp\Bigg[&\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\Theta}\mathbf{y}_i - \sum_{i=1}^{n} q(\mathbf{x}_i'\boldsymbol{\Theta}|\boldsymbol{\xi}_i) - \frac{p+1}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Psi}| \\ &- \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Psi}^{-1}\boldsymbol{\mu}\Bigg]. \end{aligned}$$

The log of the variational bound can be written

$$\begin{aligned} \log \underline{g}(\mathbf{Y}, \boldsymbol{\Theta}|\boldsymbol{\Xi}) = &-\frac{1}{2}\text{vec}\left(\boldsymbol{\Theta}\right)' \left\{\sum_{i=1}^{n} \left[\mathbf{M}(\boldsymbol{\xi}_i)^{-1} \otimes \mathbf{x}_i\mathbf{x}_i'\right] + \boldsymbol{\Psi}^{-1}\right\} \text{vec}\left(\boldsymbol{\Theta}\right) \\ &+ \left\{\sum_{i=1}^{n} \left[\mathbf{y}_i' - \nabla f(\boldsymbol{\xi}_i)' + \boldsymbol{\xi}_i'\mathbf{M}(\boldsymbol{\xi}_i)^{-1}\right] \otimes \mathbf{x}_i' + \boldsymbol{\mu}\boldsymbol{\Psi}^{-1}\right\} \text{vec}\left(\boldsymbol{\Theta}\right) - \frac{p+1}{2}\log(2\pi) \\ &- \sum_{i=1}^{n} \left\{f(\boldsymbol{\xi}_i) - \nabla f(\boldsymbol{\xi}_i)'\boldsymbol{\xi}_i + \frac{1}{2}\boldsymbol{\xi}_i'\mathbf{M}(\boldsymbol{\xi}_i)^{-1}\boldsymbol{\xi}_i\right\} - \frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Psi}^{-1}\boldsymbol{\mu}, \end{aligned} \tag{35}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$. This lower bound is proportional to a multivariate density with respect to $\boldsymbol{\Theta}$. It follows that the posterior variational approximation of $\mathrm{vec}(\boldsymbol{\Theta})|\mathbf{Y}, \boldsymbol{\Xi}$ is multivariate normal with mean $\mathrm{vec}(\boldsymbol{\eta})$ and variance $\boldsymbol{\Omega}$, where

$$\boldsymbol{\Omega}^{(t+1)} := \left\{ \sum_{i=1}^{n} \left[\mathbf{M}(\boldsymbol{\xi}_i)^{-1} \otimes \mathbf{x}_i \mathbf{x}_i' \right] + \boldsymbol{\Psi}^{-1} \right\}^{-1}, \tag{36}$$

$$\mathrm{vec}\left(\boldsymbol{\eta}^{(t+1)}\right) := \boldsymbol{\Omega}^{(t+1)} \left\{ \sum_{i=1}^{n} \left[\mathbf{y}_i' - \nabla f(\boldsymbol{\xi}_i)' + \boldsymbol{\xi}_i' \mathbf{M}(\boldsymbol{\xi}_i)^{-1} \right] \otimes \mathbf{x}_i' + \boldsymbol{\mu} \boldsymbol{\Psi}^{-1} \right\}. \tag{37}$$

To determine the matrix of the variational parameters $\boldsymbol{\Xi}$, where each $\boldsymbol{\xi}_i \in \mathbb{R}^k$, we make the variational approximation as close to the true density as possible. Because the variational approximation is a lower bound, we only need to focus on the variational parameter $\boldsymbol{\Xi}$. We follow [12] and develop a method based on the expectation-maximization (EM) algorithm [8], where $\boldsymbol{\Theta}$ is taken to be a latent variable. We follow this methodology and take the complete-data to be $(\mathbf{Y}, \boldsymbol{\Theta})$. In the E-step, we obtain the function

$$Q\left(\boldsymbol{\Xi}^{(t+1)} | \boldsymbol{\Xi}^{(t)}\right) = \mathrm{E}_{\boldsymbol{\Theta}|\mathbf{Y}, \boldsymbol{\Xi}^{(t)}} \left[\log \underline{g}(\mathbf{Y}, \boldsymbol{\Theta} | \boldsymbol{\Xi}^{(t+1)}) \right]. \tag{38}$$

Then, to minimize the difference between the true and approximating density, we minimize (38) with respect to each variational parameter $\boldsymbol{\xi}_i$.

Similarly to [12], we obtain an explicit expression for updating each $\boldsymbol{\xi}$ based on the gradient, and the derivation is given Appendix B. The updates for each row of $\boldsymbol{\xi}$ are given by

$$\boldsymbol{\xi}_i^{(t+1)} = \hat{\mathbf{z}}_i + \left(\mathbf{M}^{-1} + \mathbf{H}\right)^{-1} \left(\mathbf{w}'\mathbf{H} + \omega \mathbf{q}\right), \tag{39}$$

where

$$\begin{aligned}
\mathbf{w} &= \frac{1}{2} \mathbf{m}' \left\{ \mathrm{diag}\left(\mathbf{M}^{-1}(\hat{\mathbf{z}}_i - \boldsymbol{\xi}) \odot \mathbf{M}^{-1}(\hat{\mathbf{z}}_i - \boldsymbol{\xi}) \right) + \mathbf{M}^{-1} \hat{\mathbf{Z}}_i \mathbf{M}^{-1} \odot \mathbf{I}_k \right\}, \\
\omega &= \frac{1}{2} \left\{ \left(\mathbf{1}_k' \mathbf{M}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\xi}) \right)^2 + \mathbf{1}_k' \mathbf{M}^{-1} \hat{\mathbf{Z}}_i \mathbf{M}^{-1} \mathbf{1}_k \right\} q_{k+1} m' \left(q_{k+1}\right),
\end{aligned}$$

$\hat{\mathbf{z}}_i = \mathbf{x}_i' \boldsymbol{\eta} = \left(\mathbf{I}_k \otimes \mathbf{x}_i'\right) \boldsymbol{\eta}$, and $\hat{\mathbf{Z}}_i = \left(\mathbf{I}_k \otimes \mathbf{x}_i'\right) \boldsymbol{\Omega} \left(\mathbf{I}_k \otimes \mathbf{x}_i\right)$.

We then iterate based on equations (36), (37), and (39).

5.2. *The prevalence of pneumoconiosis among coalminers*

To illustrate our variational approximation for Bayesian multinomial regression, we consider data from [1] concerning the degree of pneumoconiosis in coalface workers as a function of exposure measured in years. The severity of disease is measured radiologically and each man is assigned to one of three classes according to the degree of abnormality revealed in his X-ray. The data are reproduced in Table 1. [15] analyze this data set in the frequentist framework using a logit model, whereas [1] use a probit model and obtain comparable results. We follow them and use the log-years exposed as the covariate because as shown in Figure 3 this leads to a linear relationship between with the degree of abnormalities and the covariate log-years.

Table 1
Period of exposure and prevalence of pneumoconiosis amongst a group of coalminers

Exposure (years)	Category I: normal	Category II	Category III: severe
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

Table 2
Posterior means and 95% credible intervals for the parameters from the Bayesian multinomial regression model applied to the coalminers data

Parameter	Mean	lower CI	upper CI
α_{II}	10.401	9.429	11.373
α_{III}	10.254	9.282	11.226
β	2.582	2.281	2.883

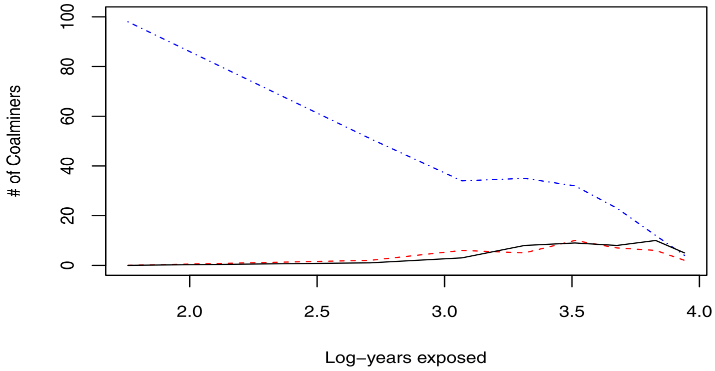

Fig 3. *The degree of pneumoconiosis versus log-years of exposure amongst a group of coalminers, where the dashed blue line is Category I (normal), the dashed red line is Category II, and the solid black line is Category III (severe).*

In our analysis, we use a multivariate normal prior for the parameters with mean equal to $\mathbf{0}_4$ and covariance matrix $1000 \times \mathbf{I}_4$. We use Category I (i.e., normal) as the reference level. Both [1] and [15] assume a common slope parameter and different intercepts. We incorporate a common slope parameter into the Bayesian multinomial regression framework by letting $\text{vec}(\boldsymbol{\Theta}) = \mathbf{P}(\alpha_{\mathrm{II}}, \alpha_{\mathrm{III}}, \beta)'$, where $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

6. An MM algorithm for multinomial regression

To assess the relative computational efficiency of the $\mathbf{M}$ bound (31) and the $\mathbf{B}$ bound (27), we compare them in the context of multinomial regression. An MM algorithm is discussed in Section 6.1 and a simulation study in then performed in Section 6.2.

6.1. The multinomial regression

In multinomial regression we maximize the likelihood function, i.e., (32), given the data. Because the estimates cannot be determined in closed form, an iteratively reweighted least squares technique is usually employed to find the maximum likelihood estimates. An alternative approach is to use the lower bound technique from [4], which fits within the class of MM algorithms [11, 14]. Herein, we derive a novel MM algorithm, based on the $\boldsymbol{\Sigma}$-strong convexity of the multivariate log-exp sum, using the methodology developed in the previous section. Using (34), we can obtain a surrogate function for the log-likelihood, i.e.,

$$\underline{l}(\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\Xi}) = \sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\Theta}\mathbf{y}_i - \sum_{i=1}^{n} q(\mathbf{x}_i'\boldsymbol{\Theta}|\boldsymbol{\xi}_i),$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)$ is a vector of expansion points. This surrogate function has an explicit solution for $\boldsymbol{\Theta}$, given by

$$\text{vec}\left(\boldsymbol{\Theta}^{\text{new}}\right) = \left[\sum_{i=1}^{n} \mathbf{M}(\boldsymbol{\xi}_i)^{-1} \otimes \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \sum_{i=1}^{n} \left[\mathbf{y}_i' - \nabla f(\boldsymbol{\xi}_i)' + \boldsymbol{\xi}_i'\mathbf{M}(\boldsymbol{\xi}_i)^{-1}\right] \otimes \mathbf{x}_i'.$$

Once we obtain a solution to the surrogate function, we construct a new surrogate function by updating the expansion points $\boldsymbol{\xi}_i$. Following work in the previous section, we can obtain updates for the expansion points as follows:

$$\boldsymbol{\xi}_i^{\text{new}} = \mathbf{x}_i'\boldsymbol{\Theta} + \frac{1}{2}(\mathbf{M}(\boldsymbol{\xi}_i)^{-1} + \mathbf{H})^{-1}\left[\mathbf{m}'\text{diag}\left(\mathbf{t}\odot\mathbf{t}\right)\mathbf{H}_i + \left[\mathbf{1}_k'\mathbf{t}\right]^2 q_{k+1}m'\left(q_{k+1}\right)\mathbf{q}_i\right],$$

where $\mathbf{t}_i = \mathbf{M}(\boldsymbol{\xi}_i)^{-1}\left(\mathbf{x}_i'\boldsymbol{\Theta} - \boldsymbol{\xi}\right)$, $\mathbf{H}_i = \text{diag}\left(\mathbf{q}_i\right) - \mathbf{q}_i\mathbf{q}_i'$, $\mathbf{q}_i = \nabla f\left(\boldsymbol{\xi}_i\right)$ and $\mathbf{m}$ is defined in Appendix B.4.2. We then iterate based on $\boldsymbol{\Theta}^{\text{new}}$ and $\boldsymbol{\xi}_i^{\text{new}}$.

6.2. Simulation comparison to the Böhning bound

To compare the computational efficiency of the the $\mathbf{M}$ bound (31) and the $\mathbf{B}$ bound (27), we simulate data from the multinomial regression model (32), using a variety of settings, and run the algorithm given in Section 6.1 using each bound. In this experiment, we vary the number of observations $n \in \{250, 500, 1000\}$ and the number of distinct categories $k \in \{2, 5\}$. The number of covariates is set equal to k. We also set $\boldsymbol{\Theta}' = [\mathbf{0}_k, \mathbf{I}_k]$, which means $\boldsymbol{\Theta}$ is a $(k+1)\times k$ matrix, the interpret terms are all zero, and only one regression co-

TABLE 3
A comparison of the average user system times, the number of iterations to convergence, and the rate, which is the time to complete one iteration, over variety of settings for n and k, each replicated 100 times

		system.time		# Iterations		Rate	
k	n	**B**	**M**	**B**	**M**	**B**	**M**
2	250	0.35	0.36	16	10	0.025	0.037
	500	0.81	0.72	22	12	0.041	0.065
	1000	0.92	1.02	13	9	0.072	0.120
3	250	1.08	0.80	51	24	0.022	0.035
	500	1.88	1.31	50	23	0.037	0.059
	1000	4.01	2.62	52	24	0.077	0.111
4	250	2.59	1.44	131	49	0.020	0.032
	500	5.80	3.51	148	65	0.039	0.057
	1000	9.98	5.18	107	42	0.093	0.126
5	250	5.48	2.56	259	87	0.021	0.031
	500	9.50	4.87	222	85	0.043	0.059
	1000	16.94	8.42	195	73	0.086	0.115

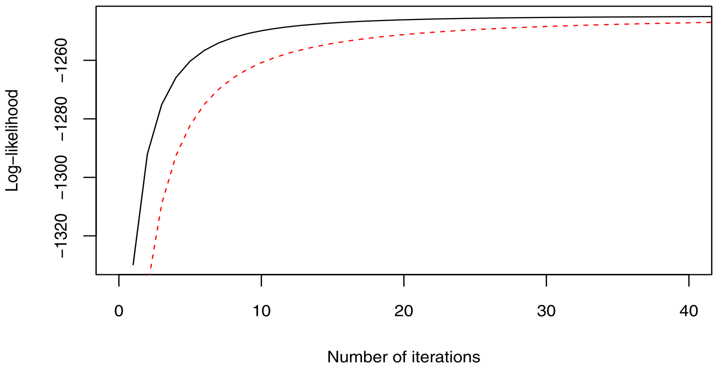

FIG 4. *Log-likelihood values versus iteration from the MM algorithm with the Böhning bound* **B** *(dashed red curve) and the* $\mathbf{M}(\mathbf{q})$ *bound (black curve), respectively.*

efficient is non-zero, i.e., $\log \pi_j = x_j$, for $j = 1, \ldots, k$. We start each algorithm using the same randomly generated starting values. In addition, we replicate each setting 100 times. In each case, we note the system times, number of iterations to convergence, and the time for one iteration for each bound. The average and standard deviation of theses quantities are given in Tables 3 and 4 (Appendix C), respectively. Figure 4 shows the results from one simulation.

The overall impression from the results indicate that the **M** bound requires more computational time for each iteration (rate) but converges much faster on average, i.e., the algorithm based on the **M** bound is slower but requires fewer iterations with the result that the overall computation time is less than for the algorithm based on the **M** bound. Table 4 (Appendix C) shows that the algorithm based on the **M** gives more consistent results in most of the experimental settings considered here.

7. Discussion

Sharp majorization has been extended to the multivariate case. The notions of σ-strong convexity, monotonicity, and one-sided Lipschitz continuity have been extended to $\mathbf{\Sigma}$-strong convexity, monotonicity, and Lipschitz continuity, respectively. We showed that these notions are interconnected and we illustrated how they connect a convex function to its Fenchel-Legendre transform. We illustrated these extensions with sharp majorization in single and multiple dimensions, and we showed that these extensions yield improvements on the bounds given in the literature.

We considered our estimation algorithm in the context of multinomial logistic regression. [4] use the Böhning bound "in the Newton-Raphson iteration instead of the Hessian matrix, leading to a monodically converging sequence of iterates." The convergence rate of the algorithm depends on $||\mathbf{I}_k - \mathbf{B}\nabla^2 L(\hat{\pi})||$ or simply how well the Hessian is approximated by the bound. We have shown that the bound given herein outperforms the bound given by [4] by developing an MM algorithm and conducting a simulation study to illustrate the computational efficiency.

From [7], we extend the logistic example to the multinomial example. One avenue for further research is to extend other examples from [7]. For instance, general logit and probit problems could be extended to general multinomial and multivariate probit, thereby extending the hinge loss function in discriminant analysis to multivariate hinge loss functions.

We develop a variational approximation for the Bayesian multinomial regression model to demonstrate one statistical application of $\mathbf{\Sigma}$-strong convexity. This approach was illustrated using data on the prevalence of pneumoconiosis among coalminers. Other variational applications will be focused on as a part of future work, and include a latent trait model for polytomous data. In the latent trait model, the manifest or observed data are categorical and the underlying latent variable is a continuous random variable typically assumed to be Gaussian. Our work in this direction will extend the work of [20], where Gauss-hermite quadrature is used to preform maximum likelihood estimation, cf. [2], but quadrature is typically used in low-dimensional setting. The methodology developed herein will lead to a variational estimation procedure and allow a higher number of factors to be used.

Appendix A: Proofs

Theorem 2.1. (1) $\Rightarrow$ (2). If we let $h(\mathbf{y}) = f(\mathbf{y}) - \frac{1}{2}\mathbf{y}'\mathbf{\Sigma}\mathbf{y}$, then we need to show h is convex given (b), i.e., we need to show

$$h((1-\alpha)\mathbf{y} + \alpha\mathbf{x}) \leqslant \alpha h(\mathbf{x}) + (1-\alpha)h(\mathbf{y}). \tag{40}$$

for some $\alpha \in (0,1)$. From (a), f is $\mathbf{\Sigma}$-strongly convex and so applying the definition given in (7) we have

$$\begin{aligned}
h((1-\alpha)\mathbf{y}+\alpha\mathbf{x}) &= f((1-\alpha)\mathbf{y}+\alpha\mathbf{x}) \\
&\qquad -\frac{1}{2}\left[(1-\alpha)\mathbf{y}+\alpha\mathbf{x}\right]'\boldsymbol{\Sigma}\left[(1-\alpha)\mathbf{y}+\alpha\mathbf{x}\right] \\
&\leqslant (1-\alpha)f(\mathbf{y})+\alpha f(\mathbf{x}) \\
&\qquad -\frac{1}{2}\alpha(1-\alpha)\left(\mathbf{y}-\mathbf{x}\right)'\boldsymbol{\Sigma}\left(\mathbf{y}-\mathbf{x}\right) \\
&\qquad\qquad -\frac{1}{2}\left[(1-\alpha)\mathbf{y}+\alpha\mathbf{x}\right]'\boldsymbol{\Sigma}\left[(1-\alpha)\mathbf{y}+\alpha\mathbf{x}\right] \\
&= (1-\alpha)f(\mathbf{y})+\alpha f(\mathbf{x}) \\
&\qquad -\frac{1}{2}\alpha(1-\alpha)\left[\mathbf{y}\boldsymbol{\Sigma}\mathbf{y}-2\mathbf{y}\boldsymbol{\Sigma}\mathbf{x}+\mathbf{x}\boldsymbol{\Sigma}\mathbf{x}\right] \\
&\qquad -\frac{1}{2}\left[(1-\alpha)^2\mathbf{y}\boldsymbol{\Sigma}\mathbf{y}+\alpha(1-\alpha)2\mathbf{y}\boldsymbol{\Sigma}\mathbf{x}+\alpha^2\mathbf{x}\boldsymbol{\Sigma}\mathbf{x}\right] \\
&= (1-\alpha)f(\mathbf{y})+\alpha f(\mathbf{x})-\frac{1}{2}(1-\alpha)\mathbf{y}\boldsymbol{\Sigma}\mathbf{y}-\frac{1}{2}\alpha\mathbf{x}\boldsymbol{\Sigma}\mathbf{x} \\
&= \alpha h(\mathbf{x})+(1-\alpha)h(\mathbf{y})
\end{aligned}$$

(2) ⇒ (3). Any proper function $g:\mathbb{R}^n\to\mathbb{R}$ is convex if and only if the mapping $\nabla g:\mathbb{R}^n\to\mathbb{R}^n$ is monotone. This implies that, for the function h, $\nabla h(\mathbf{y})=\nabla f(\mathbf{y})-\mathbf{y}'\boldsymbol{\Sigma}$ is monotone and applying the definition of monotonicity we have

$$\begin{aligned}
\left[\nabla h(\mathbf{x})-\nabla h(\mathbf{y})\right]'(\mathbf{x}-\mathbf{y}) &\geqslant 0 \\
\left[(\nabla f(\mathbf{x})-\mathbf{x}'\boldsymbol{\Sigma})-(\nabla f(\mathbf{y})-\mathbf{y}'\boldsymbol{\Sigma})\right]'(\mathbf{x}-\mathbf{y}) &\geqslant 0 \\
\left[\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\right]'(\mathbf{x}-\mathbf{y}) &\geqslant (\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),
\end{aligned}$$

which satisfies the definition of $\boldsymbol{\Sigma}$-strong mononticity.

(3) ⇒ (4). Set $\mathbf{w}_t=\mathbf{y}+t(\mathbf{x}-\mathbf{y})$ and $g(t)=f(\mathbf{w}_t)$. Then we have $g'(t)=\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y}))'(\mathbf{x}-\mathbf{y})$. Now,

$$\begin{aligned}
f(\mathbf{x})-f(\mathbf{y})=g(1)-g(0)=\int_0^1 g'(t)dt &= \int_0^1\left[\nabla f(\mathbf{w}_t)'(\mathbf{x}-\mathbf{y})\right]dt \\
&= \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y})+\int_0^1\left[\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{y})\right]'(\mathbf{x}-\mathbf{y})dt.
\end{aligned} \tag{41}$$

Therefore,

$$\left[\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{y})\right]'(\mathbf{x}-\mathbf{y})=\left[\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{y})\right]'(\mathbf{w}_t-\mathbf{y})\frac{1}{t}.$$

Because ∇f is strongly monotone,

$$\left[\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{y})\right]'(\mathbf{w}_t-\mathbf{y})\geqslant(\mathbf{w}_t-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{w}_t-\mathbf{y})=t^2(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y})$$

and so

$$\left[\nabla f(\mathbf{w}_t)-\nabla f(\mathbf{y})\right]'(\mathbf{x}-\mathbf{y})\geqslant t(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$

which we minorize to get, from (41),

$$f(\mathbf{x}) \geqslant f(\mathbf{y}) + \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}).$$

(4) ⇒ (1). Set $\mathbf{w}_t = \mathbf{y} + t(\mathbf{x}-\mathbf{y})$ for some $t \in [0,1]$. Using the inequality in (4), we have

$$f(\mathbf{x}) \geqslant f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)'(\mathbf{x}-\mathbf{y})(1-t) + \frac{1}{2}(1-t)^2(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$
$$f(\mathbf{y}) \geqslant f(\mathbf{w}_t) - \nabla f(\mathbf{w}_t)'(\mathbf{x}-\mathbf{y})\,t + \frac{1}{2}t^2(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}).$$

It follows that

$$(1-t)f(\mathbf{y}) + tf(\mathbf{x}) \geqslant f(\mathbf{w}_t) + \frac{1}{2}\left[t(1-t)^2 + (1-t)t^2\right](\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$

which can be simplified to the definition of $\boldsymbol{\Sigma}$-strong convexity. □

Theorem 2.2. (1) ⇒ (2). If we set $\mathbf{w}_t = \mathbf{y} + t(\mathbf{x}-\mathbf{y})$ then

$$f(\mathbf{x}) - f(\mathbf{y}) = \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y}) + \int_0^1 [\nabla f(\mathbf{w}_t) - \nabla f(\mathbf{y})]'(\mathbf{x}-\mathbf{y})dt. \tag{42}$$

Because ∇f is one-sided $\boldsymbol{\Sigma}$-Lipschitz continuous,

$$[\nabla f(\mathbf{w}_t) - \nabla f(\mathbf{y})]'(\mathbf{x}-\mathbf{y}) \leqslant t(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$

which we majorize to get, from (42),

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}).$$

(2) ⇒ (3). By letting $\mathbf{w}_t = \mathbf{y} + t(\mathbf{x}-\mathbf{y}) = t\mathbf{x} + (1-t)\mathbf{y}$, which means $\mathbf{x} - \mathbf{w}_t = (1-t)(\mathbf{x}-\mathbf{y})$, we obtain

$$f(\mathbf{x}) \leqslant f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)'(\mathbf{x}-\mathbf{w}_t) + \frac{1}{2}(\mathbf{x}-\mathbf{w}_t)'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{w}_t)$$
$$f(\mathbf{x}) \leqslant f(\mathbf{w}_t) + (1-t)\nabla f(\mathbf{w}_t)'(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(1-t)^2(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y})$$
$$f(\mathbf{y}) \leqslant f(\mathbf{w}_t) + t\nabla f(\mathbf{w}_t)'(\mathbf{y}-\mathbf{x}) + \frac{1}{2}t^2(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y})$$

From the latter two inequalities, it follows that

$$tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \leqslant f(\mathbf{w}_t) + \frac{1}{2}\left[t(1-t)^2 + (1-t)t^2\right](\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$
$$tf(\mathbf{y}) + (1-t)f(\mathbf{y}) \leqslant f(\mathbf{w}_t) + \frac{1}{2}t(1-t)(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}),$$

and so

$$tf(\mathbf{y}) + (1-t)f(\mathbf{y}) - \frac{1}{2}t(1-t)(\mathbf{x}-\mathbf{y})'\boldsymbol{\Sigma}(\mathbf{x}-\mathbf{y}) \leqslant f(\mathbf{w}_t).$$

(3) ⇒ (1). We can rearrange the inequality in (3) to obtain

$$f(\mathbf{x})-f(\mathbf{y})\leqslant\frac{f(\mathbf{w}_t)-f(\mathbf{y})}{t}+\frac{1}{2}(1-t)\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right).$$

If we let $t\to 0$, we have

$$f(\mathbf{x})-f(\mathbf{y})\leqslant\nabla f(\mathbf{y})'\left(\mathbf{x}-\mathbf{y}\right)+\frac{1}{2}\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right).$$

Interchanging $\mathbf{y}$ and $\mathbf{x}$, we get

$$f(\mathbf{y})-f(\mathbf{x})\leqslant\nabla f(\mathbf{x})'\left(\mathbf{y}-\mathbf{x}\right)+\frac{1}{2}\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right)$$

and adding these last two equations gives

$$\left[\nabla f(\mathbf{y})-\nabla f(\mathbf{x})\right]'\left(\mathbf{x}-\mathbf{y}\right)\leqslant\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right).$$ □

Theorem 2.3. (1) ⇒ (2). Fenchel's inequality [cf. 17] yields

$$f(\mathbf{y})+f^*(\mathbf{q})=\mathbf{q}'\mathbf{y}$$

and so $f(\mathbf{y})=\mathbf{q}'\mathbf{y}-f^*(\mathbf{q})$, for $\mathbf{q}\in\nabla f(\mathbf{y})$. If we have that $\mathbf{q}\in\nabla f(\mathbf{y})$ and $\mathbf{p}\in\nabla f(\mathbf{x})$, then

$$\mathbf{p}'\mathbf{x}-f^*(\mathbf{p})\leqslant\mathbf{q}'\mathbf{y}-f^*(\mathbf{q})+\mathbf{q}'\left(\mathbf{x}-\mathbf{y}\right)+\frac{1}{2}\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right),$$

and so

$$\left(\mathbf{p}-\mathbf{q}\right)'\mathbf{x}-\frac{1}{2}\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right)+f^*(\mathbf{q})\leqslant f^*(\mathbf{p}).$$

Now,

$$\sup_{\mathbf{x}}\left\{\left(\mathbf{p}-\mathbf{q}\right)'\mathbf{x}-\frac{1}{2}\left(\mathbf{x}-\mathbf{y}\right)'\boldsymbol{\Sigma}\left(\mathbf{x}-\mathbf{y}\right)\right\}+f^*(\mathbf{q})\leqslant f^*(\mathbf{p}).$$

The supremum is nothing but the value of conjugate of the quadratic term evaluated at $\mathbf{p}-\mathbf{q}$, i.e.,

$$\left(\mathbf{p}-\mathbf{q}\right)'\mathbf{y}+\frac{1}{2}\left(\mathbf{p}-\mathbf{q}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{p}-\mathbf{q}\right)+f^*(\mathbf{q})\leqslant f^*(\mathbf{p}),$$

which implies f^* is $\boldsymbol{\Sigma}^{-1}$ strong convex.

(2) ⇒ (1). The proof is the same as above with the inequalities in opposite directions. □

Appendix B: Variational update

If we let $\mathbf{z}_i'=\mathbf{x}_i'\boldsymbol{\Theta}$, we can simplify equation (35) to

$$\log\underline{g}(\mathbf{Y},\mathbf{X},\boldsymbol{\Theta}|\boldsymbol{\Xi})=\\ -\sum_{i=1}^{n}\left[f(\boldsymbol{\xi}_i)+\nabla f(\boldsymbol{\xi}_i)'(\mathbf{z}_i-\boldsymbol{\xi}_i)+\frac{1}{2}(\mathbf{z}_i-\boldsymbol{\xi}_i)'\mathbf{M}(\boldsymbol{\xi}_i)^{-1}(\mathbf{z}_i-\boldsymbol{\xi}_i)\right]+C,$$

where C does not depend on $\boldsymbol{\xi}_i$. Then we can write (38) as

$$Q(\boldsymbol{\xi}_i^{(t+1)}|\boldsymbol{\xi}_i) = \sum_{i=1}^{n} k(\hat{\mathbf{z}}_i, \hat{\mathbf{Z}}_i, \boldsymbol{\xi}_i) + C,$$

where the function

$$\begin{aligned} k(\hat{\mathbf{z}}_i, \hat{\mathbf{Z}}_i, \boldsymbol{\xi}_i) \quad = \quad & -f(\boldsymbol{\xi}_i) - \nabla f(\boldsymbol{\xi}_i)'(\hat{\mathbf{z}}_i - \boldsymbol{\xi}_i) \\ & - \frac{1}{2}\operatorname{tr}\left[\mathbf{M}(\boldsymbol{\xi}_i)^{-1}(\hat{\mathbf{Z}}_i + \hat{\mathbf{z}}_i\hat{\mathbf{z}}_i' - \boldsymbol{\xi}_i\mathbf{z}_i' - \mathbf{z}_i\boldsymbol{\xi}_i' + \boldsymbol{\xi}_i\boldsymbol{\xi}_i')\right], \end{aligned}$$

with

$$\begin{aligned} \hat{\mathbf{z}}_i \quad &:= \quad \mathrm{E}[\mathbf{z}_i \big| \mathbf{Y}] = \mathbf{x}_i'\mathrm{E}[\boldsymbol{\Theta}\big|\mathbf{Y}], \\ \hat{\mathbf{Z}}_i \quad &:= \quad \mathrm{Var}[\mathbf{z}_i\big|\mathbf{Y}] = (\mathbf{I}_k \otimes \mathbf{x}_i')\,\mathrm{Var}[\boldsymbol{\Theta}\big|\mathbf{Y}]\,(\mathbf{I}_k \otimes \mathbf{x}_i)\,. \end{aligned}$$

Note, the expectation and variance of $\boldsymbol{\Theta}\big|\mathbf{Y}$ are given in (37) and (36), respectively. To ease the notational burden when taking the derivative of the function k, we drop subscript i from $\hat{\mathbf{z}}_i, \hat{\mathbf{Z}}_i$, and $\boldsymbol{\xi}_i$. In addition, we let

$$k(\mathbf{z}, \mathbf{Z}, \boldsymbol{\xi}) = \sum_{j=1}^{3} w_i\left(\boldsymbol{\xi}\right),$$

with

$$\begin{aligned} w_1\left(\boldsymbol{\xi}\right) \quad &= \quad -f(\boldsymbol{\xi}) - \nabla f(\boldsymbol{\xi})'(\hat{\mathbf{z}} - \boldsymbol{\xi}), && (43) \\ w_2\left(\boldsymbol{\xi}\right) \quad &= \quad -\frac{1}{2}\operatorname{tr}\left[\mathbf{M}^{-1}(\hat{\mathbf{z}} - \boldsymbol{\xi})(\hat{\mathbf{z}} - \boldsymbol{\xi})'\right], && (44) \\ w_3\left(\boldsymbol{\xi}\right) \quad &= \quad -\frac{1}{2}\operatorname{tr}\left\{\hat{\mathbf{Z}}\mathbf{M}^{-1}\right\}. && (45) \end{aligned}$$

The derivatives of w_1, w_2, and w_3 are given in Appendices B.1, B.2, and B.3, respectively. Combining and simplifying the gradient, we obtain the expression

$$\begin{aligned} \frac{\partial k}{\partial \boldsymbol{\xi}} = (\hat{\mathbf{z}} - \boldsymbol{\xi})'\mathbf{H} + \mathbf{v} + \frac{1}{2}\mathbf{m}'\left[\operatorname{diag}\left(\mathbf{v} \odot \mathbf{v}\right) + \mathbf{V} \odot \mathbf{I}_k\right]\mathbf{H} \\ + \frac{1}{2}\left[\left(\mathbf{1}_k'\mathbf{v}\right)^2 + \mathbf{1}_k'\mathbf{V}\mathbf{1}_k\right] q_{k+1} m'\left(q_{k+1}\right)\mathbf{q}, \end{aligned}$$

where $\mathbf{v} = \mathbf{M}^{-1}(\hat{\mathbf{z}} - \boldsymbol{\xi})$ and $\mathbf{V} = \mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}$. Setting this to zero, we obtain a fixed-point algorithm for updating $\boldsymbol{\xi}$:

$$\begin{aligned} \boldsymbol{\xi}^{\text{new}} &= \hat{\mathbf{z}} + \left(\mathbf{M}^{-1} + \mathbf{H}\right)^{-1}\left\{\frac{1}{2}\mathbf{m}'\left[\operatorname{diag}\left(\mathbf{v} \odot \mathbf{v}\right) + \mathbf{V} \odot \mathbf{I}_k\right]\mathbf{H} \right. \\ & \qquad \left. + \frac{1}{2}\left[\left(\mathbf{1}_k'\mathbf{v}\right)^2 + \mathbf{1}_k'\mathbf{V}\mathbf{1}_k\right] q_{k+1} m'\left(q_{k+1}\right)\mathbf{q}\right\} \\ &= \hat{\mathbf{z}} + \left(\mathbf{M}^{-1} + \mathbf{H}\right)^{-1}\left(\mathbf{w}'\mathbf{H} + \omega\mathbf{q}\right), \end{aligned}$$

where

$$\mathbf{w} = \frac{1}{2}\mathbf{m}'\left[\mathrm{diag}\left(\mathbf{v}\odot\mathbf{v}\right) + \mathbf{V}\odot\mathbf{I}_k\right]$$

and

$$\omega = \frac{1}{2}\left[\left(\mathbf{1}_k'\mathbf{v}\right)^2 + \mathbf{1}_k'\mathbf{V}\mathbf{1}_k\right] q_{k+1} m'\left(q_{k+1}\right).$$

B.1. Gradient of w_1

$$\begin{aligned}
w_1\left(\boldsymbol{\xi}\right) &= \mathrm{tr}\left[-f(\boldsymbol{\xi}) - \nabla f(\boldsymbol{\xi})(\hat{\mathbf{z}}-\boldsymbol{\xi})'\right],\\
\mathrm{d}w_1(\mathbf{x}) &= \mathrm{tr}\left[-\mathrm{d}f(\boldsymbol{\xi}) - \mathrm{d}\nabla f(\boldsymbol{\xi})(\hat{\mathbf{z}}-\boldsymbol{\xi})' - \nabla f(\boldsymbol{\xi})\mathrm{d}(\hat{\mathbf{z}}-\boldsymbol{\xi})'\right],\\
\frac{\partial w_1}{\partial \boldsymbol{\xi}} &= -\nabla f(\boldsymbol{\xi}) + (\hat{\mathbf{z}}-\boldsymbol{\xi})' H f(\boldsymbol{\xi}) + \nabla f(\boldsymbol{\xi}) = (\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{H}.
\end{aligned}$$

B.2. Gradient of w_2

$$\begin{aligned}
w_2\left(\boldsymbol{\xi}\right) &= -\frac{1}{2}\,\mathrm{tr}\left[\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})(\hat{\mathbf{z}}-\boldsymbol{\xi})'\right],\\
\mathrm{d}w_2(\mathbf{x}) &= -\frac{1}{2}\,\mathrm{tr}\left[\mathrm{d}\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})'(\hat{\mathbf{z}}-\boldsymbol{\xi}) + \mathbf{M}^{-1}\mathrm{d}(\hat{\mathbf{z}}-\boldsymbol{\xi})(\hat{\mathbf{z}}-\boldsymbol{\xi})'\right.\\
&\qquad\left. +\,\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\mathrm{d}(\hat{\mathbf{z}}-\boldsymbol{\xi})'\right]\\
&= \frac{1}{2}\left[\left(\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\otimes\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\right)\mathrm{dvec}\mathbf{M} + 2(\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{M}^{-1}\mathrm{dvec}\boldsymbol{\xi}\right],\\
\frac{\partial w_2}{\partial \boldsymbol{\xi}} &= \frac{1}{2}\left(\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\otimes\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\right)\left[\mathbf{D}_*(\mathbf{m}) + a'\left(q_{k+1}\right)\mathbf{J}_{k^2\times k}\right]\mathbf{H}\\
&\qquad +\,(\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{M}^{-1}\\
&= \frac{1}{2}\left[(\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{M}^{-1}\odot(\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{M}^{-1}\odot\mathbf{m}\right]\mathbf{H}\\
&\qquad +\,\frac{1}{2}a'\left(q_{k+1}\right)\left[\mathbf{1}_k'\mathbf{M}^{-1}(\hat{\mathbf{z}}-\boldsymbol{\xi})\right]^2 q_{k+1}\mathbf{q} + (\hat{\mathbf{z}}-\boldsymbol{\xi})'\mathbf{M}^{-1},
\end{aligned}$$

which is obtained using the properties of $\mathbf{D}$ and the fact that $\mathbf{1}_k\mathbf{H} = q_{k+1}\mathbf{q}$.

B.3. Gradient of w_3

$$\begin{aligned}
w_3\left(\boldsymbol{\xi}\right) &= -\frac{1}{2}\,\mathrm{tr}\left\{\hat{\mathbf{Z}}\mathbf{M}^{-1}\right\},\\
\mathrm{d}w_3\left(\boldsymbol{\xi}\right) &= \frac{1}{2}\,\mathrm{tr}\left\{\hat{\mathbf{Z}}\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\mathbf{M}^{-1}\right\} = \frac{1}{2}\mathrm{vec}\left[\mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}\right]'\mathrm{dvec}\mathbf{M}\\
&= \frac{1}{2}\mathrm{vec}\left[\boldsymbol{\Lambda}'\left(\hat{\mathbf{Y}} + \hat{\mathbf{y}}\hat{\mathbf{y}}'\right)\boldsymbol{\Lambda}\right]'\left(\mathbf{M}^{-1}\otimes\mathbf{M}^{-1}\right)\mathrm{dvec}\mathbf{M}\\
&= \frac{1}{2}\mathrm{vec}\left[\mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}\right]'\mathrm{dvec}\mathbf{M},\\
\frac{\partial w_3}{\partial \boldsymbol{\xi}} &= \frac{1}{2}\mathrm{vec}\left[\mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}\right]'\left[\mathbf{D}_*(\mathbf{m}) + a'\left(q_{k+1}\right)\mathbf{J}_{k^2\times k}\right]\mathbf{H}\\
&= \frac{1}{2}\left\{\mathbf{m}'\left[\mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}\odot\mathbf{I}_k\right]\mathbf{H} + q_{k+1}a'\left(q_{k+1}\right)\mathbf{1}_k'\left[\mathbf{M}^{-1}\hat{\mathbf{Z}}\mathbf{M}^{-1}\right]\mathbf{1}_k\mathbf{q}\right\}.
\end{aligned}$$

B.4. The partial derivatives of the matrix M

The matrix $\mathbf{M}$ is defined as a function of $\mathbf{q}$ in (31), and $\mathbf{q}$ is a function $\boldsymbol{\xi}$ because from (20) we have $\mathbf{q} = \nabla f(\boldsymbol{\xi})$. To find the partial derivative of the matrix $\mathbf{M}$ with respect to $\boldsymbol{\xi}$ we use the chain rule and obtain

$$d\text{vec}\mathbf{M} = \frac{d\text{vec}\mathbf{M}}{d\mathbf{q}}\frac{d\mathbf{q}}{d\boldsymbol{\xi}}d\boldsymbol{\xi} \qquad \text{and/or} \qquad \frac{d\text{vec}\mathbf{M}}{d\boldsymbol{\xi}} = \frac{d\text{vec}\mathbf{M}}{d\mathbf{q}}\frac{d\mathbf{q}}{d\boldsymbol{\xi}}$$

For ease of notation, we define

$$\mathbf{H} := \frac{d\mathbf{q}}{d\boldsymbol{\xi}} = \nabla^2 f(\boldsymbol{\xi}) = \text{diag}(\nabla f(\boldsymbol{\xi})) - \nabla f(\boldsymbol{\xi})\nabla f(\boldsymbol{\xi}) = \text{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}'. \tag{46}$$

Note that this Hessian matrix has the property

$$\mathbf{1}_k'\mathbf{H} = \mathbf{1}_k' \,\text{diag}(\mathbf{q}) - \mathbf{1}_k\mathbf{q}\mathbf{q}' = (1 - \mathbf{1}_k'\mathbf{q})\mathbf{q}' = q_{k+1}\mathbf{q}',$$

where $q_{k+1} = (1 - \mathbf{1}_k'\mathbf{q})$.

The matrix $\mathbf{M}$ is formed by the sum of two matrices. The first matrix is a diagonal matrix given by $\mathbf{D} = \text{diag}(m(q_1), \ldots, m(q_k))$, where the function $m(q)$ is given in (31). The second matrix is a rank one matrix, denoted by $\mathbf{R}$, with the form $m(q_{k+1})\mathbf{J}_{k\times k}$ or $m(q_{k+1})\mathbf{1}_k\mathbf{1}_k'$, where $q_{k+1} = 1 - \sum_{j=1}^{k} q_j$. Therefore, we have

$$\begin{aligned}
&d\text{vec}\mathbf{M} = d\text{vec}\mathbf{D} + d\text{vec}\mathbf{R},\\
&\frac{d\text{vec}\mathbf{M}}{d\text{vec}\mathbf{x}} = \left(\frac{d\text{vec}\mathbf{D}}{d\mathbf{q}} + \frac{d\text{vec }\mathbf{R}}{d\mathbf{q}}\right)\frac{d\mathbf{q}}{d\mathbf{x}} = \left(\frac{d\text{vec}\mathbf{D}}{d\mathbf{q}} + \frac{d\text{vec }\mathbf{R}}{d\mathbf{q}}\right)\mathbf{H},\\
&d\text{vec}\mathbf{M} = d\text{vec}\mathbf{D} + d\text{vec } a(q_{k+1})\mathbf{1}_k\mathbf{1}_k' = d\text{vec}\mathbf{D} + \mathbf{1}_k \otimes \mathbf{1}_k a(q_{k+1}).
\end{aligned}$$

We derive the partial derivatives of these two matrices $\mathbf{R}$ & $\mathbf{D}$ in the Sections B.4.1 and B.4.2, respectively. Then plugging in the derivatives and simplifying we obtain

$$\frac{d\text{vec}\mathbf{M}}{d\text{vec}\mathbf{x}} = \mathbf{D}_*(\mathbf{a})\mathbf{H} - a'\left(q_{k+1}\right)\mathbf{J}_{k^2\times k}\mathbf{H} = \mathbf{D}_*\left[\mathbf{m}\left(\mathbf{I}_k \odot \mathbf{H}\right)\right] - q_{k+1}a'\left(q_{k+1}\right)\mathbf{1}_{k^2}\mathbf{q}.$$

B.4.1. The partial derivative of the rank-1 matrix R

The differential of the rank-1 matrix with respect to $\mathbf{q}$ is

$$\begin{aligned}
\frac{\partial\text{vec }\mathbf{R}}{\partial\mathbf{q}} &= (\mathbf{1}_k \otimes \mathbf{1}_k)\frac{\partial a(q_{k+1})}{\partial\mathbf{q}} = -(\mathbf{1}_k \otimes \mathbf{1}_k)a'(q_{k+1})\mathbf{1}_k'\\
&= -a'\left(q_{k+1}\right)\mathbf{1}_{k^2}\mathbf{1}_k' = -a'\left(q_{k+1}\right)\mathbf{J}_{k^2\times k}.
\end{aligned}$$

Note that, for some matrix $\mathbf{G}$,

$$\text{vec}(\mathbf{G})'\mathbf{J}_{k^2\times k} = \left(\mathbf{1}_k'\mathbf{G}\mathbf{1}_k\right)\mathbf{1}_k'.$$

B.4.2. The derivative of the diagonal matrix $\mathbf{D}$

The matrix $\mathbf{D} = \mathrm{diag}(m(q_1), \ldots, m(q_k))$ has a Jacobian equal to a matrix $\mathbf{D}_*(\mathbf{m})$, which has dimension $k^2 \times k$ with elements equal to d_i at positions $(i + k(i-1), i)$, for $i = 1, \ldots, k$, and zero otherwise. The vector $\mathbf{m}$ has elements equal to the derivatives $m(p)$ for each p_i, $i = 1, \ldots, k$. The derivatives are given by

$$m'(q) = -\frac{2(q+1)}{(1-q)^2 q} - 4\frac{\log(q)}{(1-q)^3},$$

if $q \in (0, 0.3161973762)$, and zero otherwise.

The matrix $\mathbf{D}_*\left[\mathbf{m}\right]$ has the following properties

$$\begin{aligned}\left(\mathbf{v}' \otimes \mathbf{a}'\right)\mathbf{D}_*(\mathbf{m}) &= \mathbf{v}'\mathrm{diag}(\mathbf{v} \odot \mathbf{m}) = \mathbf{v} \odot \mathbf{a} \odot \mathbf{m},\\ \mathrm{vec}\left(\mathbf{G}\right)\mathbf{D}_*(\mathbf{m}) &= \mathbf{m}'\left(\mathbf{G} \odot \mathbf{I}_k\right),\end{aligned}$$

where $\mathbf{v}$ and $\mathbf{a}$ are k dimensional column vectors and $\mathbf{G}$ is a $k \times k$ matrix.

Appendix C: Table of standard deviations

Table 4

A comparison of the standard deviation of user system times, the number of iterations to convergence, and the rate, which is the time to complete one iteration, over variety of settings for n and k, each replicated 100 times

		system.time		# Iterations		Rate	
k	n	**B**	**M**	**B**	**M**	**B**	**M**
2	250	0.51	0.33	25	11	0.0119	0.0140
	500	1.90	0.97	53	20	0.0141	0.0168
	1000	1.31	0.77	18	8	0.0049	0.0178
3	250	1.15	0.51	54	18	0.0045	0.0076
	500	3.71	1.36	99	27	0.0026	0.0066
	1000	6.93	2.60	91	26	0.0036	0.0109
4	250	3.55	1.60	181	62	0.0015	0.0039
	500	7.66	6.85	190	133	0.0008	0.0047
	1000	11.11	3.93	119	32	0.0053	0.0093
5	250	5.41	3.13	257	114	0.0006	0.0022
	500	11.62	7.82	267	141	0.0015	0.0036
	1000	20.59	13.82	236	113	0.0046	0.0067

References

[1] Ashford, J. R. (1959). An approach to the analysis of data for semiquantal respsones in biological assay. *Biometrics 15*, 573–581.

[2] Bartholomew, D. J. and Knott, M. (1999). Latent variable models and factor analysis. In *Kendall's Library of Statistics* (2nd ed.), Volume 7. London: Edward Arnold. MR1711686

[3] Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London.

[4] Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics 44*(1), 197–200. MR1165584

[5] Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics 40*(4), 641–663. MR0996690

[6] Corduneanu, A. and Bishop, C. (2001). Variational Bayesian model selection for mixture distributions. *Artificial Intelligence and Statistics 37*, 27–34.

[7] de Leeuw, J. and Lange, K. (2009). Sharp quadratic majorization in one dimension. *Computational Statistics and Data Analysis 53*(7), 2471–2484. MR2665900

[8] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 38*, 1–38. MR0501537

[9] Donchev, T. and Farkhi, E. (1998). Stability and Euler approximation of one-sided Lipschitz differential inclusions. *SIAM Journal on Control and Optimization 36*(2), 780–796. MR1616554

[10] Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In: Krzanowski, W. J. (Ed.), *Recent Advances in Descriptive Multivariate Analysis*, Volume 58. Oxford: Springer-Verlag. MR1380319

[11] Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician 58*, 30–37. MR2055509

[12] Jaakkola, T. S. W. and Jordan, M. I. W. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

[13] Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning 37*, 183–223.

[14] Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics 9*, 1–59. MR1819865

[15] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.), Volume 2. London: Chapman & Hall. MR3223057

[16] McGrory, C. A. and Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis 51*, 5352–5367. MR2370876

[17] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational Analysis*. New York: Springer-Verlag. MR1491362

[18] Subedi, S. and McNicholas, P. D. (2007). Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classifcation 8*(2), 167–193. MR3210265

[19] Teschendorff, A., Wang, Y., Barbosa-Morais, N., Brenton, J., and Caldas, C., A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics 21*.

[20] Tipping, M. (1999). Probabilistic visualisation of high-dimensional binary data. In M. Kearns, S. Solla, and D. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*, Volume 11. Cambridge, MA, USA: MIT PRESS, pp. 592–598.

[21] Waterhouse, S., MacKay, D., and Robinson, T. (1996). Bayesian methods for mixture of experts. *Advances in Neural Information Processing Systems 8*.