Variational inference based on a subclass of closed skew normals

Linda S. L. Tan
(joint work with Aoxiang Chen)

Department of Statistics & Data Science
National University of Singapore

Introduction

- Variational methods: popular, scalable option to MCMC for Bayesian inference.
- Given data y, posterior density $p(\theta|y)$ of variables θ is approximated by a variational density $q(\theta)$ (assumed to satisfy some restrictions).
- Kullback-Leibler divergence (KLD) between $q(\theta)$ and $p(\theta|y)$ is minimized. Equivalent to maximizing a lower bound on log marginal likelihood.
- **Gaussian variational approximation** is widely used.
  - Offered via automatic differentiation variational inference (ADVI) in Stan.
  - Improve normality of constrained/skewed variables via transformations.
  - **Bernstein-von Mises theorem**: posteriors in parametric models converge to Gaussian at $\mathcal{O}(1/\sqrt{n})$. Large sample sizes required for close resemblance.
- Posteriors of probit, tobit, multinomial probit models belong to unified skew normals (Anceshi et al. 2023).
- **Skewed Bernstein-von Mises theorem** (Durante et al. 2023): posteriors in parametric models converge to generalized skew normal at $\mathcal{O}(1/n)$ (faster).

Introduction

- **Former work**: Multivariate skew normal as variational density (Omerod, 2011; Lin et al., 2019; Zhou et al., 2023) and implicit copulas (Smith et al., 2020). Skew decomposable graphical models (Salomone et al., 2023).
- **Main contributions**:
  - Use as variational density a **closed skew normal (CSN)** subclass built via affine transformations of independent univariate skew normals.
  - **Flexible**: a bounding line is permitted in each dimension, unlike skew normal whose tail is bounded by a single line.
  - Show that lower bound is **stationary** when skewness parameter, $\lambda = 0$. Problems in maximum likelihood estimation of λ persist.
  - A "**centered parametrization**" (mean, transformed skewness, decomposition of covariance matrix) can resolve optimization issues.
  - Derive **analytic natural gradients** for optimizing lower bound (Cholesky or LU decomposition of covariance matrix and a data augmentation scheme).

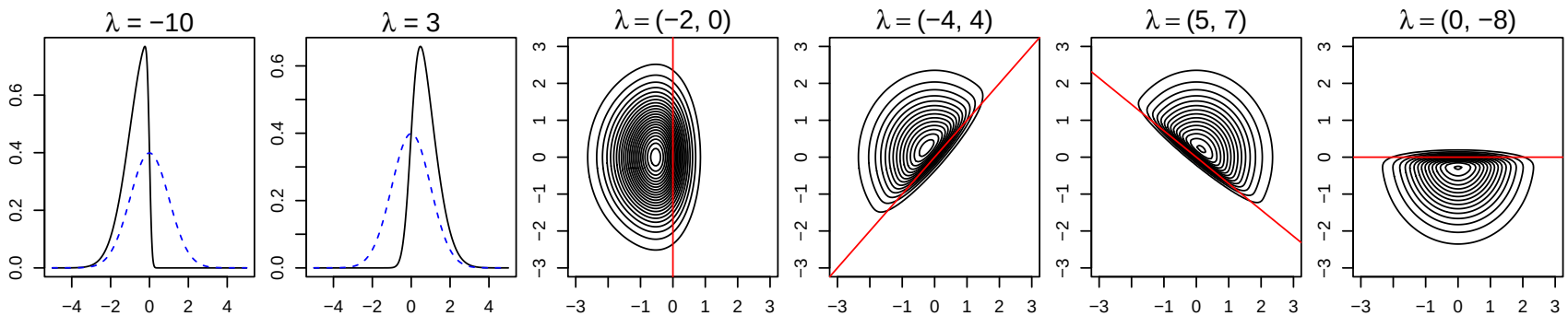
Multivariate skew normal (SN)

- If $\theta \sim \mathrm{SN}_d(\mu, \Sigma, \lambda)$,

$$p(\theta) = 2\,\phi_d(\theta|\mu, \Sigma)\,\Phi\{\lambda^\top(\theta - \mu)\}.$$

  $\mu \in \mathbb{R}^d$: location, $\lambda \in \mathbb{R}^d$: shape, Σ: $d \times d$ symmetric positive definite matrix.

- **Role of shape parameter λ** (set $\mu = 0$, $\Sigma = I_d$):
  - Univariate densities are more skewed as $|\lambda|$ increases.
  - Contour plots of bivariate densities: Any angle can be captured by varying λ, but not **densities bounded in more than one direction**.
  - This feature stems from θ being constructed as conditioned on a single random variable being positive. More limiting as dimension increases.



Closed skew normal (CSN)

- **CSN** (González-Farías et al., 2004) conditions on q random variables. If $\theta \sim \text{CSN}_{d,q}(\mu, \Sigma, D, \nu, \Delta)$,

$$p(\theta) = \phi_d(\theta|\mu, \Sigma) \frac{\Phi_q(D(\theta - \mu)|\nu, \Delta)}{\Phi_q(0|\nu, \Delta + D\Sigma D^\top)}.$$

  $\mu \in \mathbb{R}^d$, $\nu \in \mathbb{R}^q$, D: $q \times d$ matrix, Σ, Δ: $d \times d$ and $q \times q$ positive def. matrices.

- Many closure properties, due to inclusion of $\Phi_q(\cdot)$ and parameters ν and Δ, (closed under affine transformation).
- Alternative formulation of CSN is known as "unified skew normal".
- Evaluation of $\Phi_q(\cdot)$ is challenging for large q.
- Consider CSNs where $q = d$, and Δ and $\Delta + D\Sigma D^\top$ are diagonal matrices.
  - **Flexible**: capture densities with a bounding line in each dimension.
  - **Fast**: $\Phi_q(\cdot|\mu, \Sigma)$ is product of q univariate cdfs when Σ is diagonal.

  Construct **CSN subclass** by transforming independent univariate skew normals.

CSN subclass

- Let $v_i \sim \mathrm{SN}_1(0, 1, \lambda_i)$ independently for $i = 1, \ldots, d$ and $v = (v_1, \ldots, v_d)^\top$.
- $\mathrm{E}(v) = b\delta$ and $\mathrm{Var}(v) = D_\tau^2$. **Standardize** v: Define $z = D_\tau^{-1}(v - b\delta)$. Then

$$\theta = \mu + Cz,$$

  has mean μ and covariance matrix $\Sigma = CC^\top$
- As CSN is closed under **affine transformation**, $\theta \sim \mathrm{CSN}_{d,d}(\mu^*, \Sigma^*, D^*, 0, I_d)$.

  $d = 1$: reduces to SN.
  $d > 1$: does not coincide with SN.
- Transformation of θ from v: $q(\theta) = p(v)|C^{-1}||D_\tau|$.

$$\log q(\theta) = d\log(2) - \frac{d}{2}\log(2\pi) - \frac{v^\top v}{2} - \log|C| + \sum_{i=1}^{d}\{\log\Phi(\lambda_i v_i) + \log\tau_i\},$$

  where $v = D_\tau C^{-1}(\theta - \mu) + b\delta$.

Decomposition of covariance matrix

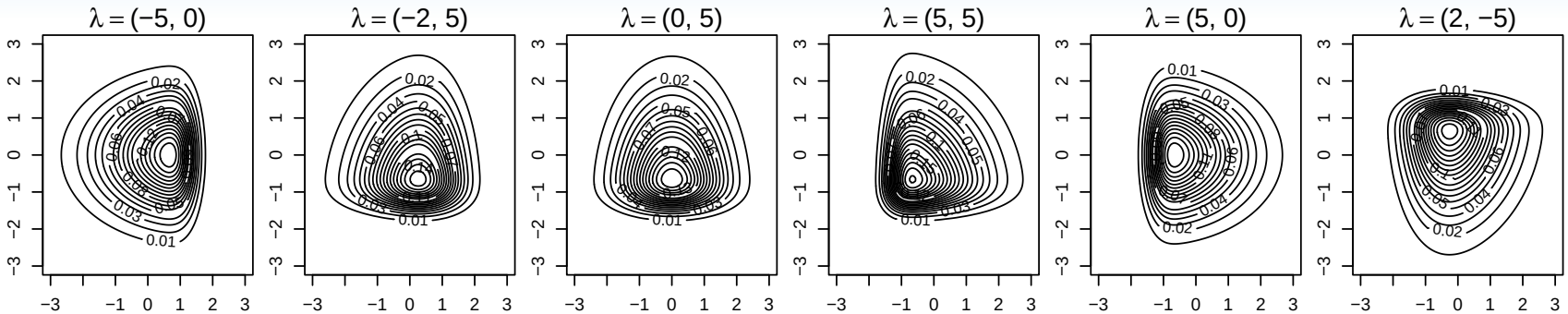


Figure 1: Bivariate density contour plots of CSN subclass ($\mu = 0,\ C = I_2$).

- Fan-shaped densities with two bounding lines can be captured by CSN subclass.
- Impossible to **rotate** densities without altering degree of flattening using λ. Role of rotation (and scaling) have fallen onto C.
- Decompose covariance matrix $\Sigma = CC^\top$: ensure Σ is positive definite and allow unconstrained optimization of lower bound.
- **Cholesky factorization**: efficient computation, (conditional) independence assumptions can be imposed easily. **Disadvantage**: C is triangular matrix (cannot capture rotations in each plane).

Decomposition of covariance matrix

- **Spectral decomposition**: Let $\Sigma = QDQ^\top$, where D is a diagonal matrix containing eigenvalues of Σ, and Q contains orthonormal eigenvectors of Σ.
  - Represent Q as $\prod_{i=1}^{d-1} \prod_{j=i+1}^{d} G_{ij}$ (Givens rotation matrices).
  - Number of parameters of $C = QD^{1/2}$ remains as $d(d+1)/2$.
  - **Disadvantage**: All scalings captured by $D^{1/2}$ must be performed *prior* to rotations by Q, limiting permissible transformations.
- As $CC^\top$ is already symmetric, we consider $C = LU$ **decomposition** (L: lower triangular, U: upper triangular with unit diagonal for uniqueness). Ensures Σ is positive definite and C^{-1} can be computed efficiently.
- Consider C as Cholesky factor (CSNC) or $C = LU$ (CSNLU).
- In high-dimensions, decomposing the covariance matrix into a factor structure can enhance computational efficiency (Ong et al. 2018). Salomone et al. (2023) use Cholesky decomposition of precision matrix to capture conditional independence structure.

Properties of CSN subclass

1. $\mathrm{E}(\theta) = \mu$ and $\mathrm{Var}(\theta) = \Sigma = CC^\top$ (by construction).
2. **Cumulant generating function** (log of moment generating function) is

$$K_\theta(t) = t^\top \mu^* + \frac{1}{2} t^\top \Sigma^* t + \sum_{j=1}^{d} \zeta_0(\alpha_j C[:, j]^\top t), \quad t \in \mathbb{R}^d.$$

3. **Marginal density** of ith element of θ is $\theta_i \sim \mathrm{CSN}_{1,d}(\mu_i^*, \Sigma_{ii}^*, D_i, 0, \Delta_i)$ and

$$q(\theta_i) = 2^d \, \phi(\theta_i | \mu_i^*, \Sigma_{ii}^*) \, \Phi_d(D_i(\theta_i - \mu_i^*) | 0, \Delta_i),$$

where $b_i = D_\tau^{-1} C[i, :]$, $\mu_i^* = \mu_i - b C[i, :]^\top \alpha$, $\Sigma_{ii}^* = b_i^\top b_i$, $D_i = D_\lambda b_i / \Sigma_{ii}^*$, $\Delta_i = I_d + D_\lambda (I_d - b_i b_i^\top / \Sigma_{ii}^*) D_\lambda$. To evaluate $\Phi_d(\cdot)$ for large d, hierarchical Cholesky factorization (Genton et al. 2018) may be useful.

Properties of CSN subclass

1. **Cumulant generating function** of θ_i is

$$K_{\theta_i}(t) = \mu_i^* t + \tfrac{1}{2}\Sigma_{ii}^* t^2 + \sum_{j=1}^{d} \zeta_0(\alpha_j C_{ij} t), \quad t \in \mathbb{R}.$$

2. **Marginal mean, variance and Pearson's index of skewness** of θ_i are μ_i, Σ_{ii} and $b(2b^2-1)(\sum_{j=1}^{d} \alpha_j^3 C_{ij}^3)/\Sigma_{ii}^{3/2}$ respectively..

3. To **simulate** from $q(\theta)$, use the stochastic representation:

$$\theta | w \sim \mathsf{N}(\mu + CD_\alpha \widetilde{w},\ CD_\kappa^2 C^\top), \quad w \sim \mathsf{N}(0, I_d),$$

where $\widetilde{w} = |w| - b\mathbf{1}$, $|w| = (|w_1|, \ldots, |w_d|)^\top$, $D_\alpha = \mathrm{diag}(\alpha)$, $D_\kappa = \mathrm{diag}(\kappa)$ and $\kappa = 1/\sqrt{1 + (1-b^2)\lambda^2}$.

Optimize evidence lower bound

- Approximate posterior $p(\theta|y)$ by variational density $q(\theta)$ with parameters η. Maximize w.r.t. η, the lower bound,

$$\mathcal{L} = \mathsf{E}_q\{h(\theta)\}, \quad \text{where} \quad h(\theta) = \log p(y, \theta) - \log q(\theta).$$

- Find optimal variational parameters by setting $\nabla_\eta \mathcal{L} = 0$.
- **Difficulties**:
  - SN reduces to Gaussian at $\lambda = 0$ and has peculiar traits in this vicinity.
  - Log-likelihood is non-quadratic, has stationary point at $\lambda = 0$, for any observed data, creating difficulties in maximum likelihood estimation.
  - Fisher information matrix is singular at $\lambda = 0$ (slow convergence, bimodal asymptotic distribution for maximum likelihood estimates).
- CSN subclass reduces to SN when $d = 1$, but our goal is to maximize lower bound. Unclear if peculiar behavior around $\lambda = 0$ will persist.
- Theorem 1 shows that the lower bound also has a stationary point at $\lambda = 0$ (unlikely to be global maximum unless posterior does not exhibit skewness).

Stationary point and alternate parametrizations

Theorem 1

Let $\hat{\mu}$ and $\hat{C}$ be optimal parameters of a Gaussian variational approximation where $\Sigma = CC^\top$. If a density $q(\theta)$ from the SN or CSN subclass is used as variational approximation instead, then the lower bound has a stationary point at $\mu = \hat{\mu}$, $C = \hat{C}$ and $\lambda = 0$.

- Optimization difficulties:
  - Iterates may get stuck at stationary point.
  - Sensitivity to initializations.
  - Slow convergence near $\lambda = 0$ due to surface flatness.
- Arellano-Valle & Azzalini (2008): Fisher information is nonsingular and log-likelihood is more quadratic-like for a **centered parametrization** based on mean, covariance and per dimension skewness.
- CSN subclass: Pearson's skewness index is $b(2b^2 - 1)\alpha^3$ in one dimension. Replace λ by α^3 or λ^3 (overcome singularities in Fisher information, Hallin & Ley, 2014).

Lower bound

- In some problems, lower bound may be evaluated in closed form or numerical integration. For CSN subclass, the **entropy**, $H_q = -\mathrm{E}_q\{\log q(\theta)\}$ is

$$\frac{d}{2}\{\log(\pi/2)+1\} + \log|C| - \sum_{i=1}^{d}[2\mathrm{E}_{\phi(u|0,1)}\{\Phi(\lambda_i u)\log\Phi(\lambda_i u)\} + \log\tau_i],$$

- H_q depends on λ and C only, has stationary point at $\lambda = 0$ and is symmetric about $\lambda = 0$. $\mathrm{E}_q\{\log p(y,\theta)\}$ is model dependent.
- Entropy plots ($C = I_2$) reveal flat region around $\lambda = 0$ (parametrize in λ). Complicates identification of optimal λ without strong data or prior influence. Many sharp corners if λ^3 is used. α^3 yields almost quadratic contours.

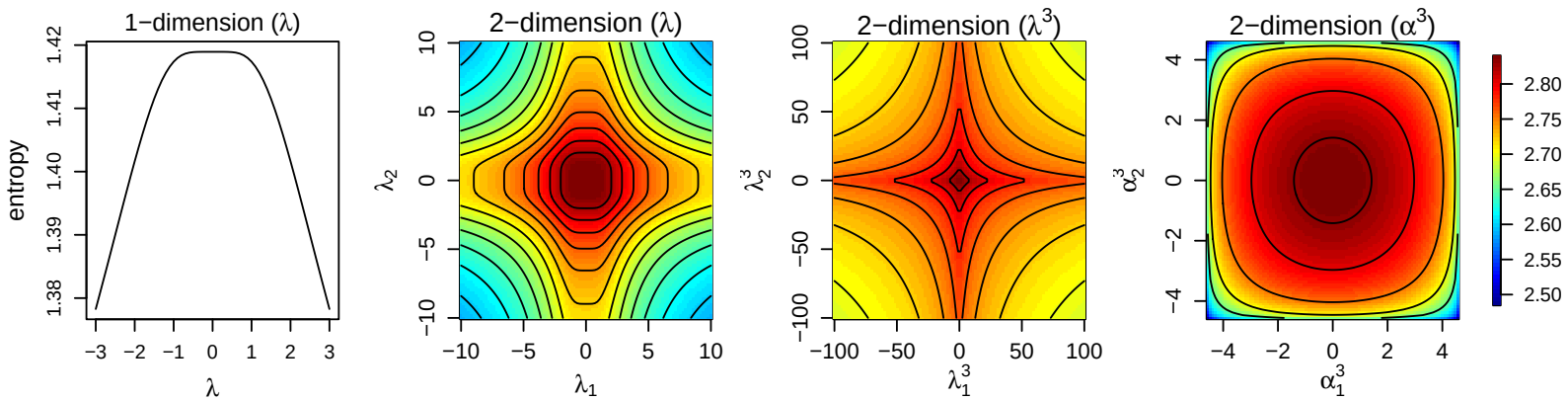


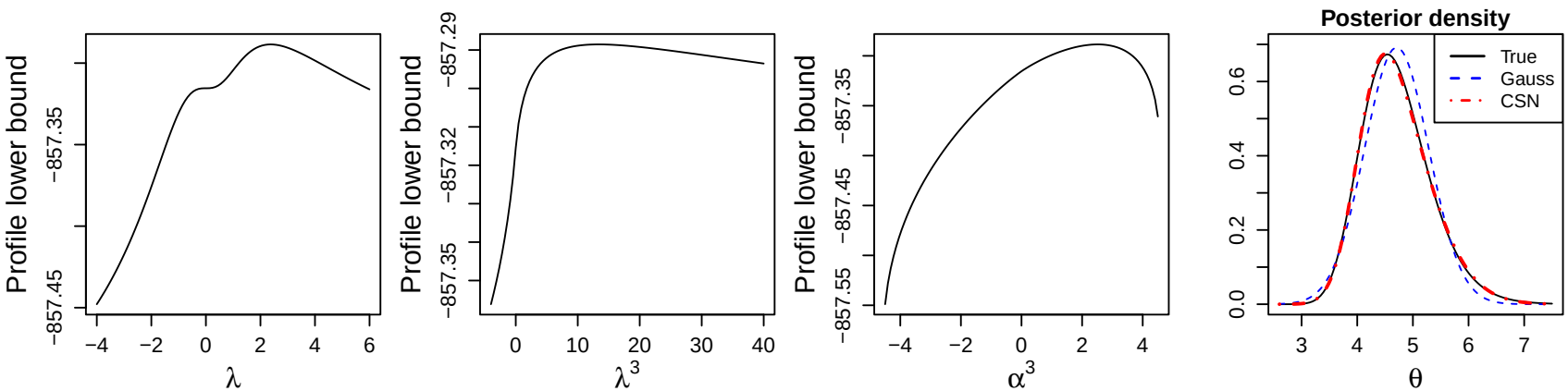
Figure 2: Plots of entropy for different parametrizations with $C = I_2$.

Accuracy assessment

- Higher $\mathcal{L}$ indicates an approximation closer in KLD to true posterior (hard to quantify how significant an improvement of 0.1 is).
- Assess accuracy via **integrated absolute error**, $\text{IAE}(q) = \int |q(\theta) - q_{\text{GS}}(\theta)| d\theta$, by comparing $q(\theta)$ with a gold standard $q_{\text{GS}}(\theta)$ (numerical integration or kernel density estimate based on MCMC samples).
- As $\text{IAE} \in (0, 2)$, we use $\text{Accuracy}(q) = 1 - \text{IAE}(q)/2$, expressed as percentage.
- MCMC (RStan): Two chains (50000 iterations each) are run in parallel. Discard first half as burn-in. Use 50000 draws left for kernel density estimate.
- **Multivariate accuracy** of variational density relative to MCMC:
  - **Cross-match non-bipartite statistic** (NBP, Yu and Bondell, 2023): Higher count of cross-match NBPs containing one sample from each distribution suggests greater similarity between the distributions.
  - **Maximum mean discrepancy** (MMD, Zhou et al., 2023): Higher $M^* = -\log(\max(\text{MMD}, 0) + 10^{-5})$ indicates better accuracy.

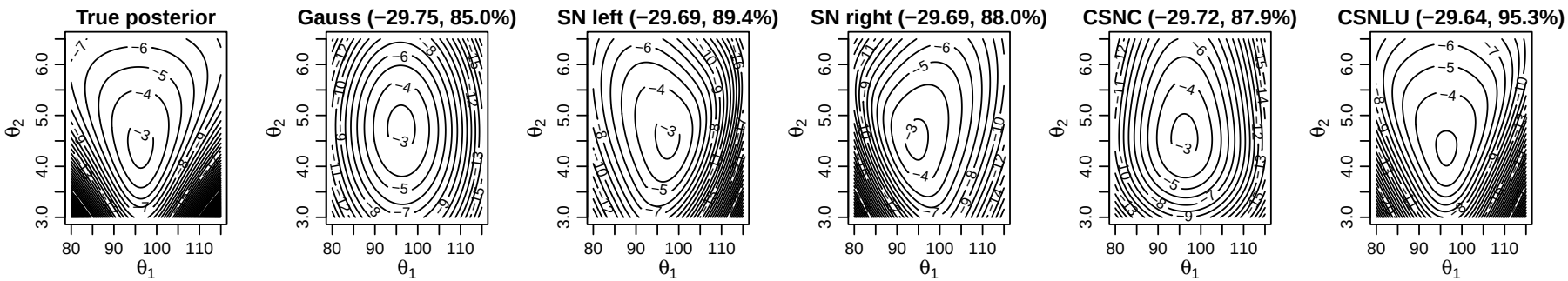
Normal sample

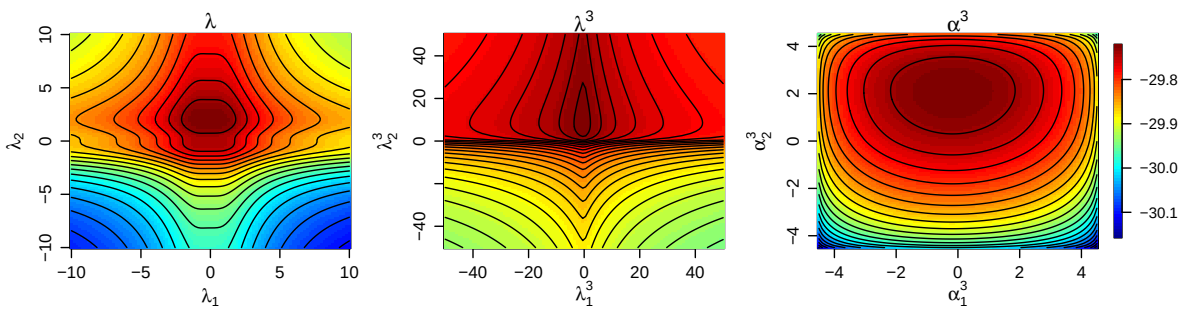
- Let $\theta = (\theta_1, \theta_2)$ and $y_i \sim \mathsf{N}(\theta_1, \exp(\theta_2))$ for $i = 1, \ldots, n$.
  Priors: $\theta_1 \sim \mathsf{N}(0, \sigma_0^2)$ and $\exp(\theta_2) \sim \mathsf{IG}(a_0, b_0)$.
- Let $\theta_1 = 0$ and $\theta = \theta_2$ only. True posterior: $\mathsf{IG}(a_0 + \frac{n}{2}, b_0 + \sum_{i=1}^n y_i^2/2)$.
- Simulate $n = 6$ observations with $\exp(\theta_2) = 225$. Write C as σ.
- **Profile lower bound**, $\mathcal{L}(\lambda) = \mathcal{L}(\hat{\mu}(\lambda), \hat{\sigma}(\lambda), \lambda)$, where $\hat{\mu}$ and $\hat{\sigma}$ maximize $\mathcal{L}$ for any given λ, is **non-quadratic** and has **stationary point** at $\lambda = 0$.
- "Centered" parameters μ and σ alone cannot eliminate stationary point, but replacing λ by λ^3 or α^3 achieves this goal (α^3 yields more pronounced mode).
- CSN has higher accuracy (99.0%) than Gaussian (92.6%).



Normal sample

- θ_1 and θ_2 both unknown: Simulate $n = 6$ observations ($\theta_1 = 100$, $\mathrm{e}^{\theta_2} = 225$).
- Compare true posterior with variational approximations computed via BFGS:
  - Gaussian is least accurate (unable to capture skewness).
  - SN can capture left or right tail depending on initialization, but not both as it is bounded in only one direction.
  - CSNC does not have correct orientation (unable to rotate).
  - CSNLU achieves highest lower bound and accuracy (more parameters).





- Lower bound of CSNLU: stationary point at $\lambda = 0$.
- α^3: quadratic-like surface.

Stochastic variational inference for CSN subclass

- $\mathrm{E}_q\{\log p(y,\theta)\}$ is not analytically tractable: maximize $\mathcal{L}$ w.r.t. η using stochastic gradient ascent.

$$t\text{th iteration:} \quad \eta_{t+1} = \eta_t + \rho_t \widehat{\nabla}_\eta \mathcal{L}(\eta_t).$$

- Compute unbiased gradient estimates via **reparametrization trick**:

$$\widehat{\nabla}_\eta \mathcal{L} = \nabla_\eta \theta \nabla_\theta h(\theta), \quad \text{where} \quad \theta = \mathcal{T}_\eta(z), z \sim p(z).$$

- **CSN subclass**:
  - Set $\theta = C(D_\kappa w_2 + D_\alpha \widetilde{w}_1) + \mu$, where $w_1, w_2 \overset{iid}{\sim} \mathsf{N}(0, I_d)$, $\widetilde{w}_1 = |w_1| - b\mathbf{1}$.
  - If C **is Cholesky factor**, then $\eta = (\mu^\top, \lambda^\top, \mathrm{vech}(C)^\top)^\top$.
  - If $C = LU$, then $\eta = (\mu^\top, \lambda^\top, \mathrm{vech}(L)^\top, \mathrm{vech}_u(U)^\top)^\top$.
  - If α^3 is used instead of λ, then

$$\widehat{\nabla}_{\alpha^3} \mathcal{L} = (\nabla_{\alpha^3} \lambda) \widehat{\nabla}_\lambda \mathcal{L}.$$

Natural gradients for CSN subclass

- Euclidean metric: not suitable when optimizing $\mathcal{L}$ in curved parameter space.
- KLD between densities: steepest ascent direction is given by **natural gradient**, which premultiplies Euclidean gradient with inverse of the Fisher information,

$$\mathcal{I}_\theta(\eta) = -\mathsf{E}_q\{\nabla_\eta^2 \log q(\theta)\}.$$

- SN: Fisher information is **singular** at $\lambda = 0$ and has **intractable expectations**.
- Derive analytic natural gradients for CSN subclass by considering Cholesky or LU decomposition, and Fisher information of $q(\theta, w)$, with **same parameter** η as $q(\theta)$. Feasible as

$$\mathcal{L} = \int q(\theta) h(\theta)\, d\theta = \int\int q(\theta, w) h(\theta)\, d\theta\, dw.$$

- Replace $\mathcal{I}_\theta(\eta)^{-1}\nabla_\eta\mathcal{L}$ by $\widetilde{\nabla}_\eta\mathcal{L} = \mathcal{I}_{\theta,w}(\eta)^{-1}\nabla_\eta\mathcal{L}$, where

$$\mathcal{I}_{\theta,w}(\eta) = \mathcal{I}_\theta(\eta) + \mathsf{E}_{q(\theta)}\{\mathcal{I}_{w|\theta}(\eta|\theta)\}.$$

- Unlike $\mathcal{I}_\theta(\eta)$ which has singularities, $\mathcal{I}_{\theta,w}(\eta)$ is **positive definite** $\rightarrow$ Derive its inverse and natural gradients analytically (more conservative progress).

Natural gradients for CSN subclass

- Let $K = \kappa^2 \mathbf{1}^\top$.

Theorem 2

For variational density $q(\theta)$ in CSN subclass, if C is Cholesky factor of Σ and $\eta = (\mu^\top, \lambda^\top, \mathrm{vech}(C)^\top)^\top$, then the natural gradient is

$$\widetilde{\nabla}_\eta \mathcal{L} = \begin{bmatrix} (CD_\kappa^2 C^\top)\nabla_\mu \mathcal{L} \\ \frac{1}{(1-b^2)(2\kappa^2-\kappa^4)} \odot \nabla_\lambda \mathcal{L} + \frac{\lambda}{2-\kappa^2} \odot \mathrm{diag}(A_1) \\ \mathrm{vech}(CA_1), \end{bmatrix},$$

where $G = \{C^\top \mathrm{vech}^{-1}(\nabla_{\mathrm{vech}(C)} \mathcal{L})\}_\ell$ and
$A_1 = \mathrm{diag}(\frac{\alpha\kappa}{2} \odot \nabla_\lambda \mathcal{L}) + G \odot \{K - \mathrm{diag}(\frac{\kappa^4}{2})\}$.

Natural gradients for CSN subclass

Theorem 3

For variational density $q(\theta)$ in CSN subclass, if $C = LU$ where L is a lower triangular matrix and U is an unit diagonal upper triangular matrix, and $\eta = (\mu^\top, \lambda^\top, \mathrm{vech}(L)^\top, \mathrm{vech}_u(U)^\top)^\top$, then the natural gradient is

$$
\begin{aligned}
\widetilde{\nabla}_\mu \mathcal{L} &= (C D_\kappa^2 C^\top) \nabla_\mu \mathcal{L}, \\
\widetilde{\nabla}_{\mathrm{vech}(L)} \mathcal{L} &= \mathrm{vech}(L \mathcal{G}_\ell), \\
\widetilde{\nabla}_\lambda \mathcal{L} &= \tfrac{1}{(1-b^2)(2\kappa^2-\kappa^4)} \odot \nabla_\lambda \mathcal{L} + \tfrac{\lambda}{2-\kappa^2} \odot \mathrm{diag}(H), \\
\widetilde{\nabla}_{\mathrm{vech}_u(U)} \mathcal{L} &= \mathrm{vech}_u[U\{K_u \odot (F - H^\top)\} + \mathcal{G}_u U],
\end{aligned}
$$

where $a = \mathrm{vech}\{\mathrm{diag}(\tfrac{1}{2-\kappa^2}) + (1/K)_\ell - (K_u)^\top\}$,

$$
\begin{aligned}
G &= \{L^\top \mathrm{vech}^{-1}(\nabla_{\mathrm{vech}(L)} \mathcal{L})\}_\ell, \qquad \mathcal{G} = UHU^{-1}, \\
F &= \{U^\top \mathrm{vech}_u^{-1}(\nabla_{\mathrm{vech}_u(U)} \mathcal{L})\}_u, \quad A_2 = \mathrm{vech}^{-1}(1/a),
\end{aligned}
$$

$$
H = A_2 \odot [U^\top \{G - (U^{-T} F U^\top)_\ell\} U^{-T} + \mathrm{diag}(\tfrac{\lambda}{2-\kappa^2} \odot \nabla_\lambda \mathcal{L}) - (K \odot F)^\top].
$$

Applications

- Logistic regression, survival models, zero-inflated negative binomial models and generalized linear mixed models.
- Compare CSN variational approximations with planar and real NVP normalizing flows using Gaussian as a baseline.
- **Logistic regression**

  Binomial responses: Bioassay data ($n = 4,\ d = 2$)
  Binary responses: German credit data ($n = 1000,\ d = 49$)

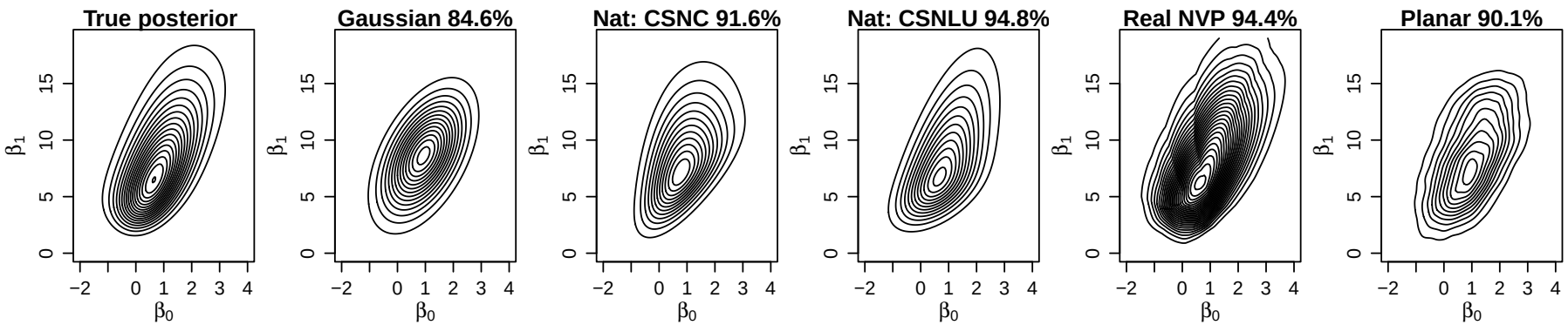
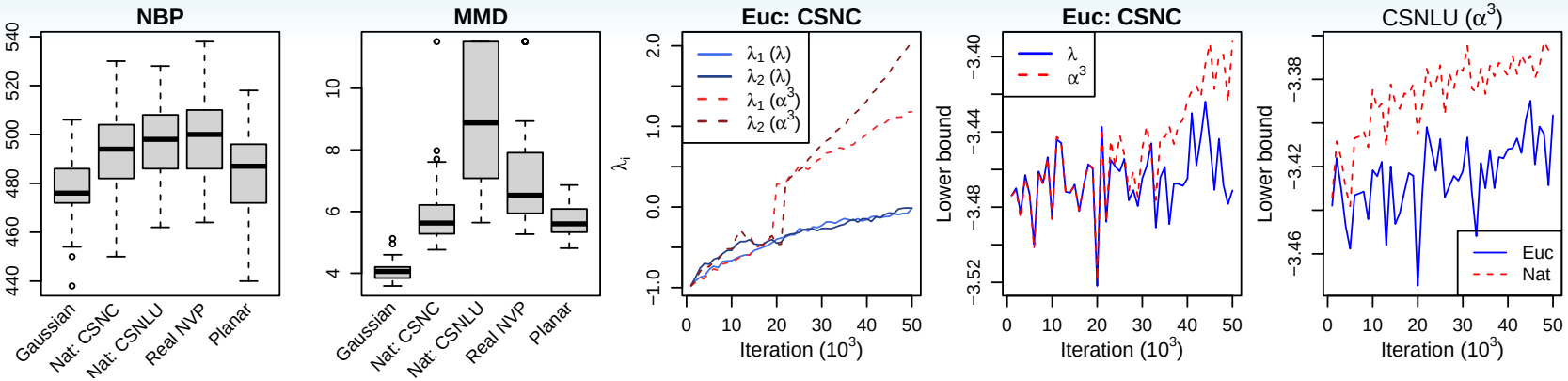


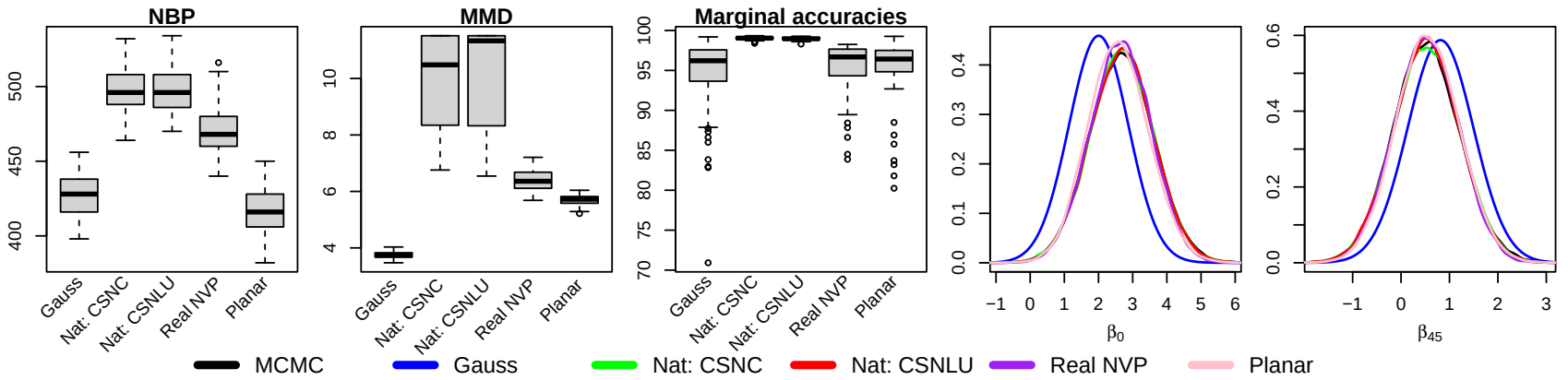
Figure 3: Bioassay data: Contour plots of true posterior and variational approximations

- CSNLU and real NVP flow have highest accuracies of 94–95%.
- Orientation of CSNC does not match true posterior well (inability to rotate).

Logistic regression



- λ converges slowly and is stuck at zero (λ parametrization), but iterates can escape zero stationary point and converge to better mode (α^3 parametrization).
- Natural gradients yield higher lower bounds than Euclidean gradients.



- CSNC and CSNLU outperform Gaussian, real NVP and planar flows in NBP, MMD and marginal density estimates.

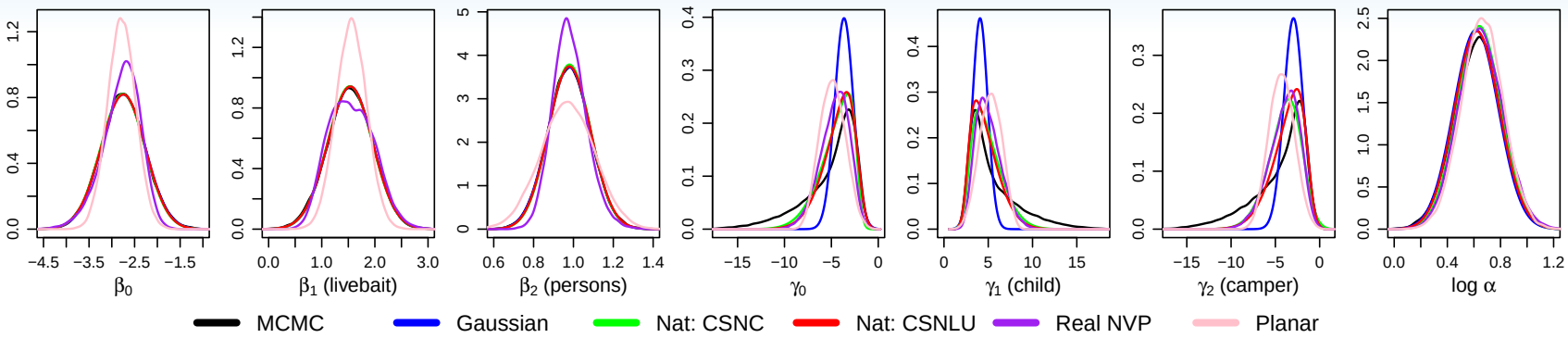
Zero-inflated negative binomial model

- Fish dataset: number of fish (y_i) caught by visitor i in a day for $i = 1, \ldots, 250$.
- y_i is 0 with probability φ_i, or is generated from a negative binomial distribution (which has parameters α and μ_i) with probability $1 - \varphi_i$.
- $\log \mu_i = x_i^\top \beta$ and $\text{logit}(\varphi_i) = z_i^\top \gamma$. Thus $\theta = (\beta^\top, \gamma^\top, \log \alpha)^\top$ and $d = 7$.
- CSNLU provides best fit based on lower bound, NBP and MMD, while CSNC is second best.
- For marginal density estimates, Gaussian is highly accurate for β (∼99%), but less so for γ (∼67%) whose marginal posteriors are highly asymmetrical.

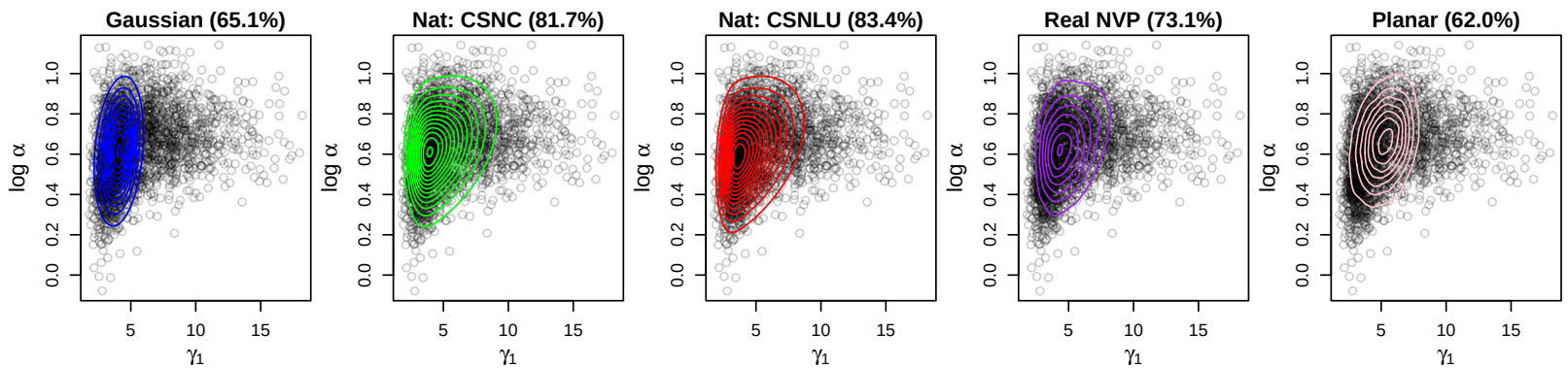
	$\mathcal{L}$	NBP	MMD	β_0	β_1	β_2	γ_0	γ_1	γ_2	$\log \alpha$
Gaussian	-425.8	401.2	2.4	99.0	99.1	99.4	67.4	65.5	68.1	95.0
Nat: CSNC	-425.3	461.0	3.6	99.0	98.8	98.8	83.9	83.2	78.1	96.9
Nat: CSNLU	**-425.2**	**473.5**	**3.9**	99.0	99.1	99.2	84.7	85.1	85.2	96.5
Real NVP	-425.9	369.6	3.0	90.4	93.9	89.1	75.3	73.8	77.3	95.8
Planar	-426.2	330.4	2.1	78.7	81.5	88.2	64.3	62.2	62.3	92.2

Table 1: Average lower bound, NBP and MMD, and accuracies of marginal densities.

Zero-inflated negative binomial model



- CSN improves on Gaussian, and CSNLU (∼85%) slightly outperforms CSNC (∼83%). Real NVP (∼75%) also improves on Gaussian, but not as well as CSN, while planar flow is weakest.
- Bivariate marginal posterior of $(\gamma_1, \log \alpha)$ is shaped irregularly. Contour plots of planar flow and Gaussian are elliptically shaped (inadequate), while real NVP, CSNC and CSNLU can capture skewness in tail more effectively.



Generalized linear mixed model

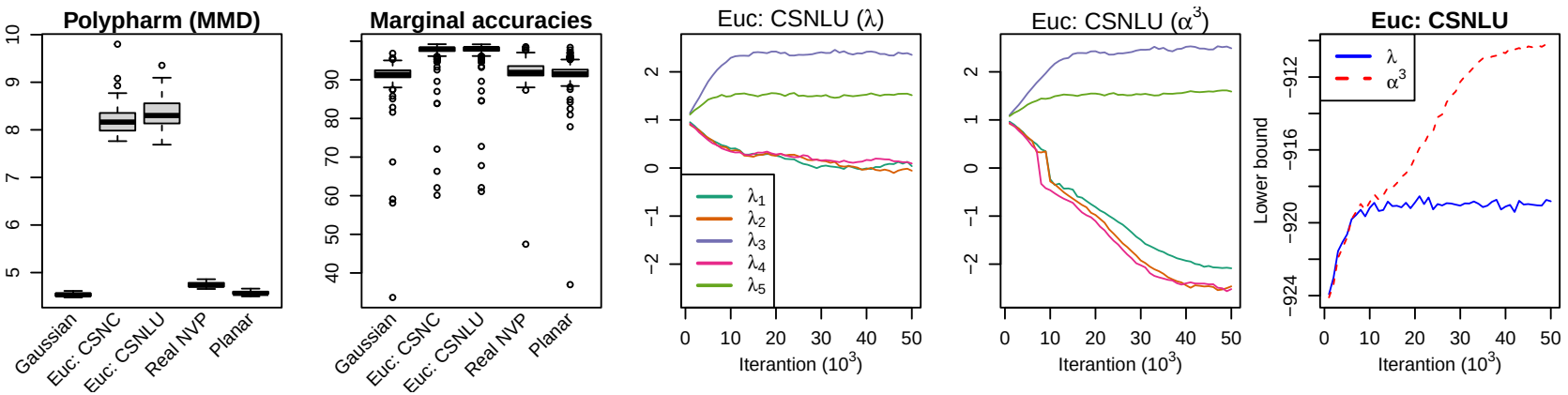
- Observe $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ for $i = 1, \ldots, n$ and let $\mu_{ij} = \mathsf{E}(y_{ij})$.

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i \quad \text{for} \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i.$$

  where β denote fixed effects, $b_i \sim \mathsf{N}(0, G^{-1})$ are random effects.

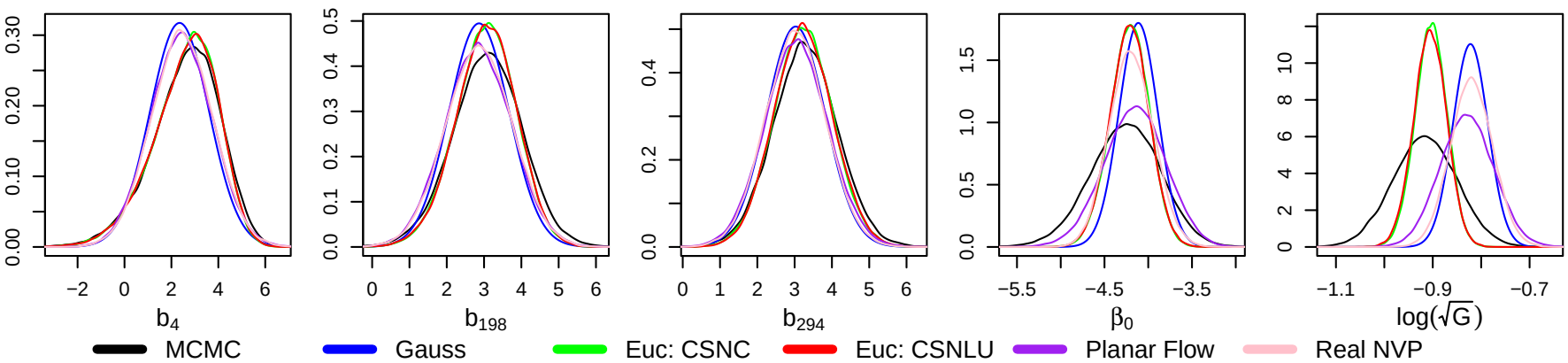
- Mean-field variational approximation: $q(\theta) = q(\theta_G) \prod_{i=1}^n q(b_i)$.

- Logistic random intercept model for polypharm dataset (Hosmer et al., 2013): 7 binary responses for each of 500 subjects observed for drug usage.



- CSNC and CSNLU provide approximations of joint and marginal posteriors closest to MCMC. Iterates of first five elements of λ are unable to traverse stationary point at zero under λ parametrization. The α^3 parametrization resolves this issue and achieves a much higher lower bound.

Generalized linear mixed model



- Marginal density estimates of some variables whose Gaussian approximation accuracy $< 90\%$.
- CSN often captures posterior modes and skewness accurately, especially for random effects, but posterior variance of global variables are underestimated due to mean-field assumption.
- Real NVP provides better estimates of the posterior variance but modal estimates are slightly misaligned.

Computation time

- Gaussian and CSN algorithms are run in Julia, MCMC in RStan, and normalizing flows in Python on the GPU.
- CSN algorithms provide speedup relative to MCMC and are often faster than flow-based methods.
- CSNC and CSNLU have similar runtimes with Euclidean gradients, but natural gradients increases their runtimes by a larger margin as dimension increases. For CSNLU, computation of natural gradients is more intensive due to inversion of unit upper triangular matrices.

	Gaussian	Euc: CSNC	Nat: CSNC	Euc: CSNLU	Nat: CSNLU	NVP	Planar	MCMC
Bioassay	0.1	0.9	1.6	1.0	4.4	242.5	199.0	45.8
German	2.2	5.4	11.6	9.7	60.6	245.6	217.2	511.2
Fish	1.1	3.2	4.0	3.3	8.1	169.6	159.4	377.9
Hip	0.3	1.4	2.2	1.5	5.2	279.1	210.1	85.2
Polypharm	9.8	22.6	29.1	23.5	50.4	469.2	244.1	723.1
Simulated	51.0	209.1	323.0	272.1	488.9	4809.9	298.0	2664.6
Diabetics	254.9	1023.3	1315.1	1292.3	1794.0	-	4470.2	13462.3

Table 2: Runtimes in seconds for all applications.

Conclusion

- Introduce closed skew normal (CSN) subclass as alternative to Gaussian variational approximation (accommodate skewness, flexible, bounding line is permitted in each dimension unlike skew normal).
- Construct subclass using affine transformations and highlight limitations in constraining linear map to be a lower triangular matrix. LU decomposition is proposed to ensure ease in inversion during optimization.
- Prove that a stationary point at zero skewness exists in maximizing variational lower bound, which creates problems in stochastic gradient ascent. Parametrizing in terms of α^3 can resolve these issues
- Derive analytic natural gradients for maximizing lower bound. Achieved by considering the augmentation $q(\theta, w)$, and Cholesky or LU decomposition.
- Investigate proposed methods using various statistical applications.