# Bayesian learning of sparse factor loadings

Magnus Rattray
School of Computer Science, University of Manchester

Bayesian Research Kitchen, Ambleside, September 6th 2008

# Talk Outline

- Brief overview of popular sparsity priors
- Example application: sparse Bayesian factor analysis for network inference
- Theory for average-case performance with mixture prior
- Theory and MCMC results for different data distributions
- Comparison with L1 prior
- Discussion

# Popular sparsity priors

Sparsity priors tend to be convenient rather than realistic, e.g.

# Popular sparsity priors

Sparsity priors tend to be convenient rather than realistic, e.g.

- L1

$$p(w_i) = \frac{\lambda}{2} e^{-\lambda |w_i|}$$

# Popular sparsity priors

Sparsity priors tend to be convenient rather than realistic, e.g.

- L1

$$p(w_i) = \frac{\lambda}{2} e^{-\lambda |w_i|}$$

- ARD

$$p(w_i) = \mathcal{N}(w_i | 0, \lambda_i^{-1})$$

# Popular sparsity priors

Sparsity priors tend to be convenient rather than realistic, e.g.

- L1

$$p(w_i) = \frac{\lambda}{2} e^{-\lambda |w_i|}$$

- ARD

$$p(w_i) = \mathcal{N}(w_i | 0, \lambda_i^{-1})$$

- Mixture

$$p(w_i) = (1 - C)\delta(w_i) + C\mathcal{N}(w_i | 0, \lambda^{-1})$$

# Popular sparsity priors

Sparsity priors tend to be convenient rather than realistic, e.g.

- L1

$$p(w_i) = \frac{\lambda}{2} e^{-\lambda |w_i|}$$

- ARD

$$p(w_i) = \mathcal{N}(w_i | 0, \lambda_i^{-1})$$

- Mixture

$$p(w_i) = (1 - C)\delta(w_i) + C\mathcal{N}(w_i | 0, \lambda^{-1})$$

Is it OK to not worry about whether these actually fit the data?

# Example application: factor analysis for network inference

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{WZ} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

$$\begin{aligned}
\mathbf{Y} &= [y_{in}] && \text{log-expression of gene } i \text{ in sample } n \\
\mathbf{Z} &= [z_{jn}] && \text{log-concentration (or "activity") of TF } j \text{ in sample } n \\
\mathbf{W} &= [w_{ij}] && \text{factor loading is "effect" of TF } j \text{ on gene } i
\end{aligned}$$

- Model $\mathbf{Z}$ as latent variable, since mRNA data may not capture TF protein level/activity, or TFs too weakly expressed
- For e.g. yeast $\mathbf{W}$ roughly $6000 \times 200$
- Various methods for inferring sparse $\mathbf{W}$ from $\mathbf{Y}$ (reviewed by Pournara and Wernisch, BMC Bioinformatics 2007).

# Factor analysis for network inference

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{WZ} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- Mixture prior leads to tractable Gibbs sampler

$$p(w_{ij}) = (1 - C_{ij})\delta(w_{ij}) + C_{ij}\mathcal{N}(w_{ij}|0, \lambda^{-1})$$

- Hyper-parameters $C_{ij} \in [0, 1]$ can be obtained from e.g.
  - ChIP-chip data
  - DNA motifs (Sabatti and James, Bioinformatics 2006)
- Or we can estimate (grouped) hyper-parameters by MCMC

# Gibbs sampler

Write $w_{ij} = x_{ij} b_{ij}$ where $x_{ij} \in \{0, 1\}$ and $b_{ij} \sim \mathcal{N}(0, \lambda^{-1})$

$$
\begin{aligned}
x_{\cdot j} &\sim p(x_{\cdot j} | \mathbf{X} \setminus x_{\cdot j}, \mathbf{Z}, \mathbf{Y}) \quad (1) \\
\mathbf{B} &\sim p(\mathbf{B} | \mathbf{X}, \mathbf{Z}, \mathbf{Y}) \quad (2) \\
\mathbf{Z} &\sim p(\mathbf{Z} | \mathbf{X}, \mathbf{B}, \mathbf{Y}) \quad (3)
\end{aligned}
$$

# Gibbs sampler

Write $w_{ij} = x_{ij} b_{ij}$ where $x_{ij} \in \{0, 1\}$ and $b_{ij} \sim \mathcal{N}(0, \lambda^{-1})$

$$
\begin{aligned}
x_{\cdot j} &\sim p(x_{\cdot j} | \mathbf{X} \setminus x_{\cdot j}, \mathbf{Z}, \mathbf{Y}) \quad (1) \\
\mathbf{B} &\sim p(\mathbf{B} | \mathbf{X}, \mathbf{Z}, \mathbf{Y}) \quad (2) \\
\mathbf{Z} &\sim p(\mathbf{Z} | \mathbf{X}, \mathbf{B}, \mathbf{Y}) \quad (3)
\end{aligned}
$$

- Integrate out $\boldsymbol{B}$ before sampling $\boldsymbol{X}$
- (2,3) more efficient when $\mathbf{X}$ is typically sparse
- Can also sample hyper-parameters $C_{ij}$ and $\lambda$ if required

# Average case theory for sparse Bayesian PCA

$$
\begin{aligned}
\boldsymbol{y}_n &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^T) \\
p(\boldsymbol{w}|C, \lambda) &= \prod_{i=1}^{N} \left[(1 - C)\delta(w_i) + C\mathcal{N}(w_i|0, \lambda^{-1})\right]
\end{aligned}
$$

# Average case theory for sparse Bayesian PCA

$$
\begin{aligned}
\boldsymbol{y}_n &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^T) \\
p(\boldsymbol{w}|C,\lambda) &= \prod_{i=1}^{N} \left[(1-C)\delta(w_i) + C\mathcal{N}(w_i|0,\lambda^{-1})\right]
\end{aligned}
$$

- We study average behaviour over datasets $D = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ produced by a teacher distribution

# Average case theory for sparse Bayesian PCA

$$
\begin{aligned}
\boldsymbol{y}_n &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^T) \\
p(\boldsymbol{w}|C,\lambda) &= \prod_{i=1}^{N} \left[(1-C)\delta(w_i) + C\mathcal{N}(w_i|0,\lambda^{-1})\right]
\end{aligned}
$$

- We study average behaviour over datasets $D = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ produced by a teacher distribution
- The teacher is identical except for a different factorized parameter distribution. We consider two cases:

# Average case theory for sparse Bayesian PCA

$$\begin{aligned} \boldsymbol{y}_n &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^T) \\ p(\boldsymbol{w}|C,\lambda) &= \prod_{i=1}^{N} \left[(1-C)\delta(w_i) + C\mathcal{N}(w_i|0,\lambda^{-1})\right] \end{aligned}$$

- We study average behaviour over datasets $D = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ produced by a teacher distribution
- The teacher is identical except for a different factorized parameter distribution. We consider two cases:

(1) Same form: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$

# Average case theory for sparse Bayesian PCA

$$\begin{aligned} \boldsymbol{y}_n &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{w}\boldsymbol{w}^T) \\ p(\boldsymbol{w}|C,\lambda) &= \prod_{i=1}^{N} \left[(1-C)\delta(w_i) + C\mathcal{N}(w_i|0,\lambda^{-1})\right] \end{aligned}$$

- We study average behaviour over datasets $D = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ produced by a teacher distribution
- The teacher is identical except for a different factorized parameter distribution. We consider two cases:

(1) Same form: $p(w_i^t) = (1-C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0,\lambda_t^{-1})$

(2) Different form: $p(w_i^t) = (1-C_t)\delta(w_i^t) + C_t\delta(w_i^t - \lambda_t^{-1/2})$

## Average case theory for sparse Bayesian PCA

- Compute marginal likelihood $\langle \log p(D|C,\lambda)\rangle_D$ in limit $N \rightarrow \infty$ with $\alpha = M/N$ held constant using replica method

# Average case theory for sparse Bayesian PCA

- Compute marginal likelihood $\langle \log p(D|C,\lambda)\rangle_D$ in limit $N \to \infty$ with $\alpha = M/N$ held constant using replica method
- Compute functions of the mean posterior parameter $\boldsymbol{w}^*(D)$

$$
\begin{aligned}
\rho(\boldsymbol{w}^*) &= \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}^t}{||\boldsymbol{w}^*||||\boldsymbol{w}^t||} \\
\mathcal{L}(\boldsymbol{w}^*) &= \langle \log p(\boldsymbol{y}|\boldsymbol{w}^*, C, \lambda)\rangle_{\boldsymbol{y}|\boldsymbol{w}^t}
\end{aligned}
$$

# Average case theory for sparse Bayesian PCA

- Compute marginal likelihood $\langle \log p(D|C,\lambda) \rangle_D$ in limit $N \to \infty$ with $\alpha = M/N$ held constant using replica method
- Compute functions of the mean posterior parameter $\boldsymbol{w}^*(D)$

$$\begin{aligned} \rho(\boldsymbol{w}^*) &= \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}^t}{||\boldsymbol{w}^*||||\boldsymbol{w}^t||} \\ \mathcal{L}(\boldsymbol{w}^*) &= \langle \log p(\boldsymbol{y}|\boldsymbol{w}^*, C, \lambda) \rangle_{\boldsymbol{y}|\boldsymbol{w}^t} \end{aligned}$$

- Good agreement with simulations for most relevant case of small $\alpha$ (so-called large $N$ small $p$ regime)

# Average case theory for sparse Bayesian PCA

- Compute marginal likelihood $\langle \log p(D|C,\lambda)\rangle_D$ in limit $N \to \infty$ with $\alpha = M/N$ held constant using replica method
- Compute functions of the mean posterior parameter $\boldsymbol{w}^*(D)$

$$\begin{aligned} \rho(\boldsymbol{w}^*) &= \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}^t}{||\boldsymbol{w}^*||||\boldsymbol{w}^t||} \\ \mathcal{L}(\boldsymbol{w}^*) &= \langle \log p(\boldsymbol{y}|\boldsymbol{w}^*, C, \lambda)\rangle_{\boldsymbol{y}|\boldsymbol{w}^t} \end{aligned}$$

- Good agreement with simulations for most relevant case of small $\alpha$ (so-called large $N$ small $p$ regime)

# Average case theory for sparse Bayesian PCA

- Similar replica calculation to Uda and Kabashima (J. Phys. Soc. Japan 74, 2005)

$$Z(D) = p(D|C,\lambda) = \int \mathrm{d}\boldsymbol{w}\, p(\boldsymbol{w}|C,\lambda) \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{w})$$

$$\begin{aligned}\frac{1}{N}\langle \log Z(D)\rangle_{D,\boldsymbol{w}^t} &= \frac{1}{N}\lim_{n\to 0}\frac{\partial}{\partial n}\langle Z^n(D)\rangle_{D,\boldsymbol{w}^t} \\ &= \alpha\langle \log p(\boldsymbol{y}|\boldsymbol{w}^*(D), C, \lambda)\rangle_{\boldsymbol{y},D,\boldsymbol{w}^t} + \text{entropic terms}\end{aligned}$$

# Average case theory for sparse Bayesian PCA

- Similar replica calculation to Uda and Kabashima (J. Phys. Soc. Japan 74, 2005)

$$Z(D) = p(D|C,\lambda) = \int \mathrm{d}\boldsymbol{w}\, p(\boldsymbol{w}|C,\lambda) \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{w})$$

$$\begin{aligned} \frac{1}{N}\langle \log Z(D)\rangle_{D,\boldsymbol{w}^t} &= \frac{1}{N}\lim_{n\to 0}\frac{\partial}{\partial n}\langle Z^n(D)\rangle_{D,\boldsymbol{w}^t} \\ &= \alpha\langle \log p(\boldsymbol{y}|\boldsymbol{w}^*(D),C,\lambda)\rangle_{\boldsymbol{y},D,\boldsymbol{w}^t} + \text{entropic terms} \end{aligned}$$

- Average case becomes typical for large $N$ due to self-averaging

# Standard PCA result (C=1)

Learning exhibits phase transitions, e.g. (for $\alpha > 1$)

$$\rho(w^*) = \theta(\alpha - T^{-2})\,\theta\!\left(\alpha - \frac{\lambda}{NT}\right)\sqrt{\frac{\alpha - T^{-2}}{\alpha + T^{-1}}}$$

where $\theta(x)$ is the step function and

$$T = ||w_t||^2_{N\to\infty} = NC_t\lambda_t^{-1} \ .$$

# Standard PCA result (C=1)

Learning exhibits phase transitions, e.g. (for $\alpha > 1$)

$$\rho(w^*) = \theta(\alpha - T^{-2})\,\theta\!\left(\alpha - \frac{\lambda}{NT}\right)\sqrt{\frac{\alpha - T^{-2}}{\alpha + T^{-1}}}$$

where $\theta(x)$ is the step function and

$$T = ||w_t||^2_{N\to\infty} = NC_t\lambda_t^{-1} \ .$$

- Consistent with result for Bayesian PCA with spherical prior $p(\boldsymbol{w}) \propto \delta(||\boldsymbol{w}|| - 1)$ (Riemann et al. J. Phys. A 1996)

# Standard PCA result (C=1)

Learning exhibits phase transitions, e.g. (for $\alpha > 1$)

$$\rho(w^*) = \theta(\alpha - T^{-2})\, \theta\!\left(\alpha - \frac{\lambda}{NT}\right) \sqrt{\frac{\alpha - T^{-2}}{\alpha + T^{-1}}}$$

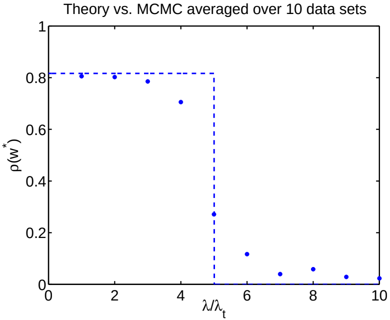
where $\theta(x)$ is the step function and

$$T = ||w_t||^2_{N \to \infty} = N C_t \lambda_t^{-1} \ .$$

- Consistent with result for Bayesian PCA with spherical prior $p(\boldsymbol{w}) \propto \delta(||\boldsymbol{w}|| - 1)$ (Riemann et al. J. Phys. A 1996)
- Only new feature is 1st-order transition with increasing $\lambda$

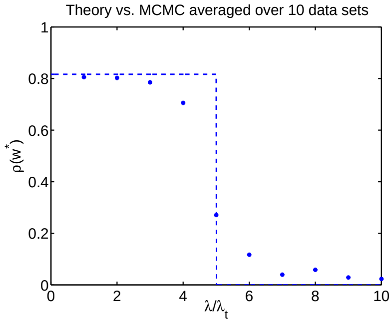
# Standard PCA result (C=1)

$$\lambda_t = N, N = 5000, M = 20000\ (\alpha = 5)$$



# Standard PCA result (C=1)

$\lambda_t = N, N = 5000, M = 20000\ (\alpha = 5)$




Here we will only consider learning away from phase transitions

# Results for sparse PCA (C<1)

- We consider two types of data set distribution:

# Results for sparse PCA (C<1)

- We consider two types of data set distribution:

(1) Same form as prior: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$

# Results for sparse PCA (C<1)

- We consider two types of data set distribution:

(1) Same form as prior: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t \mathcal{N}(w_i^t | 0, \lambda_t^{-1})$

(2) Different form: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t \delta(w_i^t - \lambda_t^{-1/2})$

# Results for sparse PCA (C<1)

- We consider two types of data set distribution:

(1) Same form as prior: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$

(2) Different form: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\delta(w_i^t - \lambda_t^{-1/2})$

- Both give identical performance for standard PCA

# Results for sparse PCA (C<1)

- We consider two types of data set distribution:

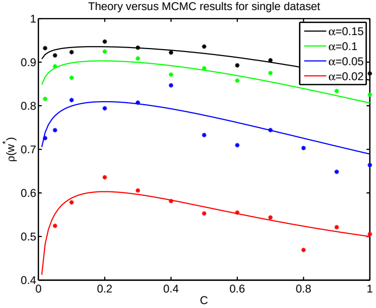
(1) Same form as prior: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$

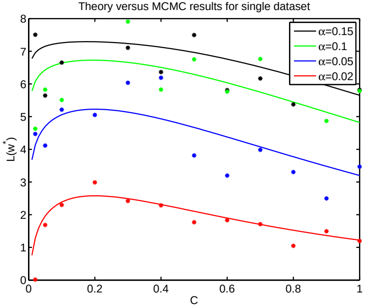
(2) Different form: $p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\delta(w_i^t - \lambda_t^{-1/2})$

- Both give identical performance for standard PCA
- Both give identical performance if sparsity is known

# Results for data distribution (1): $\rho(w^*)$ and $\mathcal{L}(w^*)$

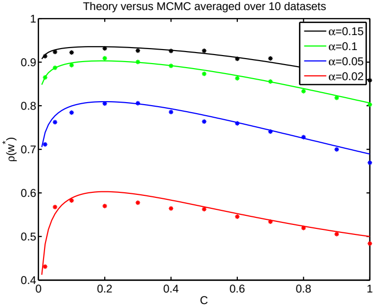
$$p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$$
$$C_t = 0.2, \lambda = \lambda_t = N/100, M = 200, N = M/\alpha$$

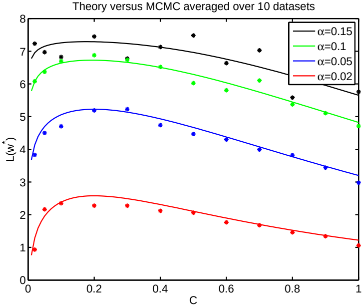


# Results for data distribution (1): $\rho(w^*)$ and $\mathcal{L}(w^*)$

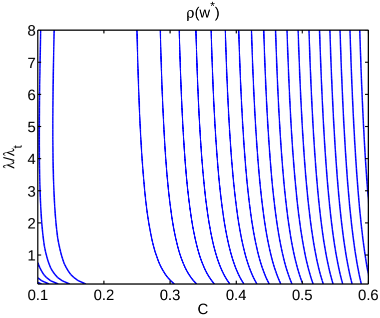
$$p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\mathcal{N}(w_i^t|0, \lambda_t^{-1})$$
$$C_t = 0.2, \lambda = \lambda_t = N/100, M = 200, N = M/\alpha$$

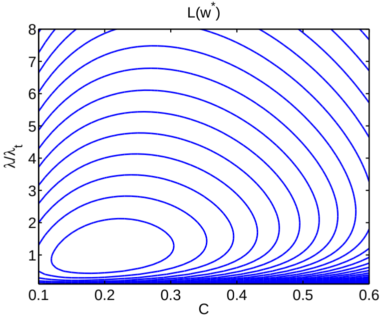


# Results for data distribution (1): $\rho(w^*)$ and $\mathcal{L}(w^*)$

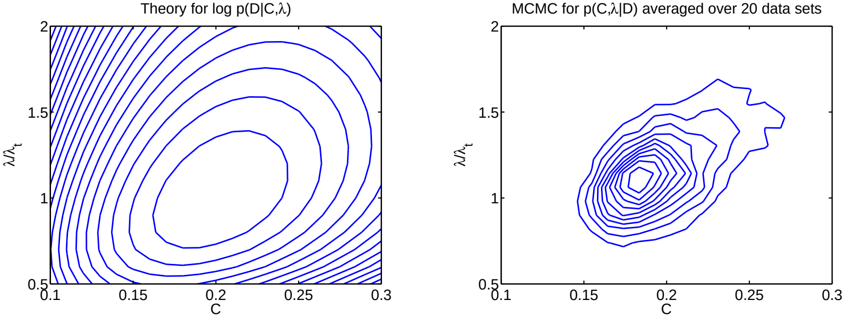
$$C_t = 0.2, \lambda_t = 20, M = 200, N = 2000\ (\alpha = 0.1)$$



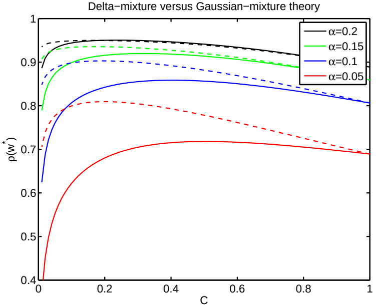
# Results for data distribution (1): $p(D|C, \lambda)$

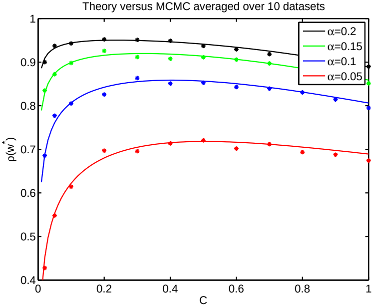
$$C_t = 0.2, \lambda_t = 20, M = 200, N = 2000\ (\alpha = 0.1)$$



# Results for data distribution (2): $\rho(w^*)$

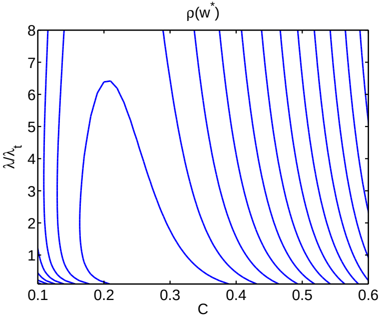
$$p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t\delta(w_i^t - \lambda_t^{-1/2})$$
$$C_t = 0.2, \lambda = \lambda_t = N/100, M = 200, N = M/\alpha$$

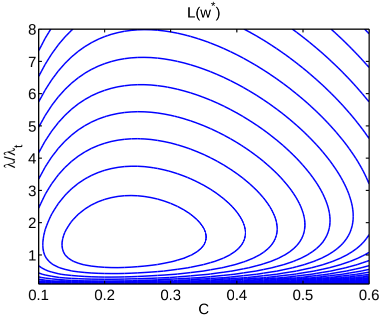


# Results for data distribution (2): $\rho(w^*)$ and $\mathcal{L}(w^*)$

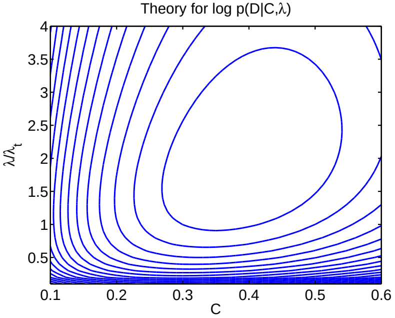
$$C_t = 0.2, \lambda_t = 10, M = 200, N = 1000 \; (\alpha = 0.2)$$

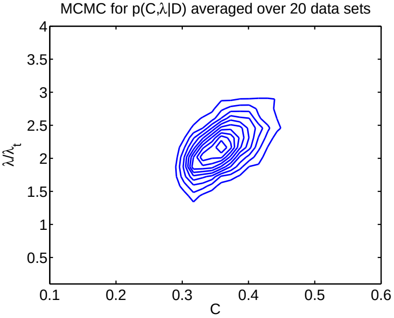


# Results for data distribution (2): $p(D|C,\lambda)$

$$C_t = 0.2, \lambda_t = 10, M = 200, N = 1000 \ (\alpha = 0.2)$$



# Other priors?

- Is this problem specific to the mixture prior?

# Other priors?

- Is this problem specific to the mixture prior?
- Consider the L1 prior (with an additional L2 term),

$$p(w_i) \propto e^{-\frac{\lambda_2 w_i^2}{2} - \lambda_1 |w_i|}$$

# Other priors?

- Is this problem specific to the mixture prior?
- Consider the L1 prior (with an additional L2 term),

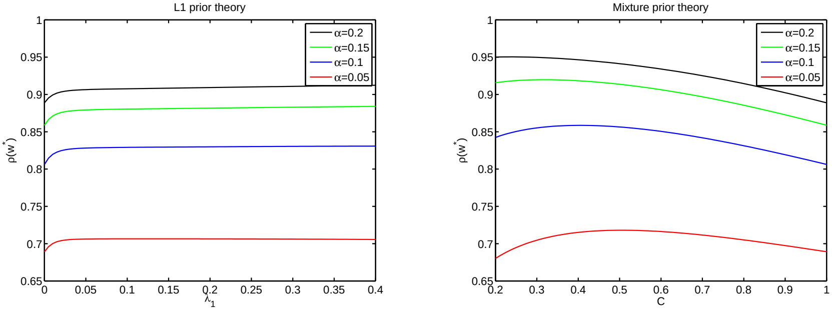
$$p(w_i) \propto e^{-\frac{\lambda_2 w_i^2}{2} - \lambda_1 |w_i|}$$

- Data distribution (2)

$$p(w_i^t) = (1 - C_t)\delta(w_i^t) + C_t \delta(w_i^t - \lambda_t^{-1/2})$$

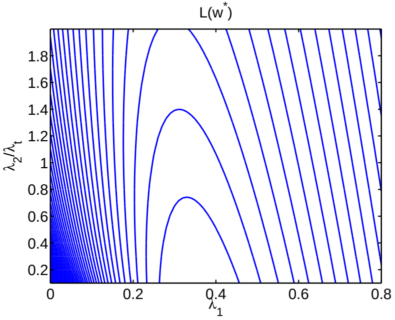
# L1 prior results: $\rho(w^*)$

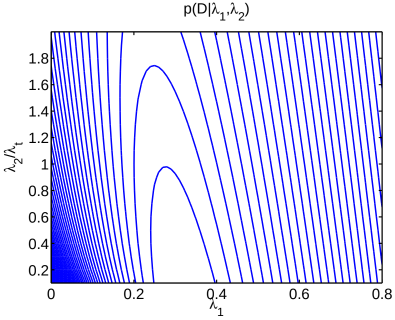
$$C_t = 0.2, \lambda_t = N/100, \alpha = M/N$$



# L1 prior results: $p(D|\lambda_1, \lambda_2)$ versus $\mathcal{L}(w^*)$ and $\rho(w^*)$

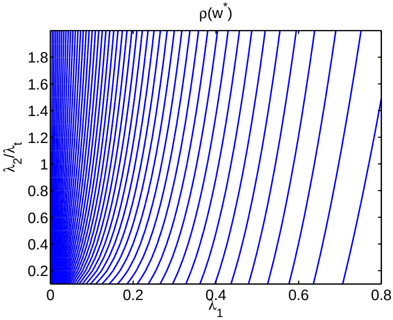
$$C_t = 0.2, \lambda_t = N/100, \alpha = 0.2$$

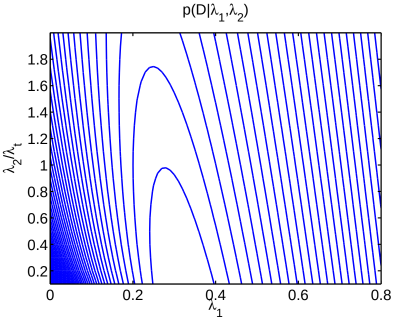


# L1 prior results: $p(D|\lambda_1, \lambda_2)$ versus $\mathcal{L}(w^*)$ and $\rho(w^*)$

$C_t = 0.2, \lambda_t = N/100, \alpha = 0.2$



# Discussion

- Mixture prior works as expected when well-matched to data

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior
- ... although L1 didn't really perform well (preliminary)

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior
- . . . although L1 didn't really perform well (preliminary)
- Future work should look at multiple factors

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior
- . . . although L1 didn't really perform well (preliminary)
- Future work should look at multiple factors
- Assessment of metrics using the full posterior

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior
- . . . although L1 didn't really perform well (preliminary)
- Future work should look at multiple factors
- Assessment of metrics using the full posterior
- Comparison with MAP and ML approaches

# Discussion

- Mixture prior works as expected when well-matched to data
- Marginal likelihood for mixture prior can be misleading
- Marginal likelihood seems more effective for L1 prior
- . . . although L1 didn't really perform well (preliminary)
- Future work should look at multiple factors
- Assessment of metrics using the full posterior
- Comparison with MAP and ML approaches
- And better priors!