1 Appendix

$(A+UC^{-1}V)^{-1} = A^{-1} - A^{-1}U(C+VA^{-1}U)^{-1}VA^{-1}$.

Hoeffiding (1) $\mathbb{E}[\exp(sX)] \leq \exp\left(s^2(b-a)^2/8\right)$ (2) $X_i \in [a_i.b_i], S_n = \sum_{i=1}^n X_i, \mathbf{P}\{S_n - \mathbb{E}_X S_n \leqslant t\} \leq \exp\left(-2t^2/\sum_{i=1}^n (b_i - a_i)^2\right)$

$\frac{\partial}{\partial \Sigma} \log|\Sigma| = \Sigma^{-T}$.

$\mathcal{N}(\mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x-\mu))$.

$R = \log \mathcal{N}(\mu, \Sigma) = -\frac{1}{2}|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$, $\frac{\partial}{\partial \mu} R = \Sigma^{-1}(x-\mu)$, $\frac{\partial}{\partial \Sigma^{-1}} R = \frac{1}{2}\Sigma - \frac{1}{2}(x-\mu)(x-\mu)^\top$.

Gaussian conditional: $\mathbb{E}[y_2|y_1] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1)$, $\mathrm{Cov}[y_2 \mid y_1] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

2 Anomaly Detection

EM

$\log p_\theta(x) = E_z[\log \frac{p_\theta(x,z)}{q(z)}] + E_z[\log \frac{q(z)}{p_\theta(z|x)}] = M(\theta, q) + E(\theta, q)$ $\max_{\theta'} M(\theta', q^*) \geq \log p_\theta(x)$

E-step: $q^* = \mathbf{argmin}_q\, E(\theta^{t-1}, q)$

M-step: $\theta^t = \mathbf{argmax}_\theta\, M(\theta, q^*)$

PCA Projection

$Var(u^\top x) = u^\top S u$, $S = (x-\overline{x})(x-\overline{x})^\top$.

3 Regression

Bias-Variance trade-off

$\mathcal{D}$ training dataset, $\hat{f}$ predictive function. $\mathbb{E}_D \mathbb{E}_{Y|X}(\hat{f}(X) - Y)^2 = \mathbb{E}_D(\hat{f}(x) - \mathbb{E}_D \hat{f}(x))^2 + \left(\mathbb{E}_D \hat{f}(x) - \mathbb{E}(Y \mid X)\right)^2 + \mathbb{E}_D(\mathbb{E}(Y \mid X) - Y)^2 = \text{Model Variance} + \text{Bias}^2 + \text{Intrinsic Noise}$.

Regularization

Ridge and Lasso can be viewed as MAP estimation with a prior on β. Ridge = Gaussian Prior and LASSO = Laplacian prior. Using SVD, we get Ridge has built-in model selection: $X\beta^{\text{Ridge}} = \sum_{j=1}^d [d_j^2/(d_j^2 + \lambda)] u_j u_j^T Y$ (each $u_j u_j^T Y$ can be viewed as a model). Lasso has more sparse estimations because the gradient of regularization does not shrink as Ridge.