Some figures and materials courtesy David Barber,
Bayesian Reasoning and Machine Learning
http://www.cs.ucl.ac.uk/staff/d.barber/brml/
BP for Continuous Variables?

**Integral-Product is Not an Algorithm!**

Is there a finitely parameterized, closed form for the message and marginal functions?

Is there an analytic formula for the message integral, phrased as an update of these parameters?

\[
p(x_1) \propto \int \int \psi_1(x_1) \psi_{12}(x_1, x_2) \psi_2(x_2) \psi_{23}(x_2, x_3) \psi_3(x_3) \psi_{24}(x_2, x_4) \psi_4(x_4) \, dx_1 \, dx_2
\]

\[
\propto \psi_1(x_1) \int \int \psi_{12}(x_1, x_2) \psi_2(x_2) \psi_{23}(x_2, x_3) \psi_3(x_3) \psi_{24}(x_2, x_4) \psi_4(x_4) \, dx_2 \, dx_3 \, dx_2
\]

\[
\propto \psi_1(x_1) \int \psi_{12}(x_1, x_2) \psi_2(x_2) \left[ \int \psi_{23}(x_2, x_3) \psi_3(x_3) \psi_{24}(x_2, x_4) \psi_4(x_4) \, dx_3 \right] \, dx_2
\]

\[
\propto \psi_1(x_1) \int \psi_{12}(x_1, x_2) \psi_2(x_2) \left[ \int \psi_{23}(x_2, x_3) \psi_3(x_3) \, dx_3 \right] \cdot \left[ \int \psi_{24}(x_2, x_4) \psi_4(x_4) \, dx_4 \right] \, dx_2
\]

\[
m_{21}(x_1) \propto \int \psi_{12}(x_1, x_2) \psi_2(x_2) m_{32}(x_2) m_{42}(x_2) \, dx_2
\]
Multivariate Gaussian Distributions
**Covariance & Correlation**

**Covariance:**
\[
\text{cov} \left[ X, Y \right] \triangleq \mathbb{E} \left[ (X - \mathbb{E} [X])(Y - \mathbb{E} [Y]) \right] = \mathbb{E} [XY] - \mathbb{E} [X] \mathbb{E} [Y]
\]

\[
\text{cov} \left[ x \right] \triangleq \mathbb{E} \left[ (x - \mathbb{E} [x])(x - \mathbb{E} [x])^T \right] = \begin{pmatrix}
\text{var} [x_1] & \text{cov} [x_1, x_2] & \cdots & \text{cov} [x_1, x_d] \\
\text{cov} [x_2, x_1] & \text{var} [x_2] & \cdots & \text{cov} [x_2, x_d] \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov} [x_d, x_1] & \text{cov} [x_d, x_2] & \cdots & \text{var} [x_d]
\end{pmatrix}
\]

- **Always** positive semidefinite: \( u^T \Sigma u \geq 0 \) for any \( u \in \mathbb{R}^{d \times 1}, u \neq 0 \)
- **Often** positive definite: \( u^T \Sigma u > 0 \) for any \( u \in \mathbb{R}^{d \times 1}, u \neq 0 \)

**Correlation:**
\[
\text{corr} \left[ X, Y \right] \triangleq \frac{\text{cov} [X, Y]}{\sqrt{\text{var} [X] \text{var} [Y]}} \quad -1 \leq \text{corr} \left[ X, Y \right] \leq 1
\]

**Independence:**
\[
p(X, Y) = p(X)p(Y) \quad \rightarrow \quad \text{cov} \left[ X, Y \right] = 0 \quad \leftrightarrow \quad \text{corr} \left[ X, Y \right] = 0
\]
Gaussian Distributions

• Probability density above assumes covariance is positive definite
• Simplest joint distribution that can capture arbitrary mean & covariance. From exponential family theory, it maximizes entropy for a given covariance.
• Arises in average of independent observations: central limit theorem
• Exponential family: ML parameters are sample mean & sample covariance

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \]

\[ \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
Two-Dimensional Multivariate Gaussians

- **Spherical**
- **Diagonal**
- **Full**
Intuition:

\[ p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \propto \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) + \frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right\} \]
Partitioned Gaussian Distributions

\[ p(x) = \mathcal{N}(x|\mu, \Sigma) \]

\[ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \]

**Marginals:**

\[ p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11}) \]

\[ p(x_2) = \mathcal{N}(x_2|\mu_2, \Sigma_{22}) \]

**Conditionals:**

\[ p(x_1|x_2) = \mathcal{N}(x_1|x_1|2, \Sigma_{1|2}) \]

\[ \mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \]

\[ = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (x_2 - \mu_2) \]

\[ = \Sigma_{1|2} (\Lambda_{11} \mu_1 - \Lambda_{12} (x_2 - \mu_2)) \]

\[ \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \Lambda_{11}^{-1} \]}
Gaussian Conditionals & Marginals

\[ p(x_1) = \mathcal{N}(x_1 | \mu_1, \sigma_1^2) \]

\[ p(x_1 | x_2) = \mathcal{N} \left( x_1 | \mu_1 + \frac{\rho \sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1^2 (1 - \rho^2) \right) \]
Directed Gaussian Graphical Models
Linear Gaussian Systems

\[ p(x) = \mathcal{N}(x | \mu_x, \Sigma_x) \quad p(y | x) = \mathcal{N}(y | Ax + b, \Sigma_y) \]

Marginal Likelihood:

\[ p(y) = \mathcal{N}(y | A\mu_x + b, \Sigma_y + A\Sigma_x A^T) \]

Posterior Distribution:

\[ p(x | y) = \mathcal{N}(x | \mu_{x|y}, \Sigma_{x|y}) \]

\[ \Sigma_{x|y} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A \]

\[ \mu_{x|y} = \Sigma_{x|y} \left[ A^T \Sigma_y^{-1} (y - b) + \Sigma_x^{-1} \mu_x \right] \]

\[ \mu_{x|y} = \mu_x + \Sigma_x A^T (A\Sigma_x A^T + \Sigma_y)^{-1} (y - A\mu_x - b) \]

\[ \Sigma_{x|y} = \Sigma_x - \Sigma_x A^T (A\Sigma_x A^T + \Sigma_y)^{-1} A\Sigma_x \]
Directed Gaussian Graphical Models

\[ \mathcal{N}(x \mid 0, \Sigma) = \prod_{s \in \mathcal{V}} \mathcal{N}(x_s \mid A_s x_{\Gamma(s)}, R_s) \]

• Sequence of locally normalized conditional distributions of each D-dimensional node:

\[
\begin{bmatrix}
\mu_s \\
\hline
A_s \\
\hline
R_s
\end{bmatrix}
\]

• Easy to modify model to capture arbitrary mean:

\[ \mathcal{N}(x \mid \mu, \Sigma) = \prod_{s \in \mathcal{V}} \mathcal{N}(x_s \mid \mu_s + A_s (x_{\Gamma(s)} - \mu_{\Gamma(s)}), R_s) \]

• Often assume zero mean, since easy to handle non-zero mean by subtracting mean from inputs, adding mean to outputs. *The challenge is in the covariance structure.*
Probabilistic PCA & Factor Analysis

- **Both Models:** Data mean is a linear function of latent coordinates

\[
p(x_i \mid z_i, \theta) = \mathcal{N}(x_i \mid Wz_i + \mu, \Psi)
\]
\[
p(z_i \mid \theta) = \mathcal{N}(z_i \mid 0, I)
\]
\[
p(x_i \mid \theta) = \mathcal{N}(x_i \mid \mu, WW^T + \Psi)
\]

- **Factor analysis:** $\Psi$ is a general diagonal matrix
- **Probabilistic PCA:** $\Psi = \sigma^2 I$ is a multiple of identity matrix

Chris Bishop, *Pattern Recognition & Machine Learning*
Linear State Space Models

Discrete Variables

Continuous Variables

Independent Observations

Sequential Observations

Mixture Models & K-means

Hidden Markov Model

Factor Analysis & PCA

State Space Model

Discrete Variables

Continuous Variables

Independent Observations

Sequential Observations
Linear (Gaussian) State Space Models

\[
\begin{align*}
    p(z, x) &= p(z)p(x | z) = p(z_1) \prod_{t=2}^{T} p(z_t | z_{t-1}) \cdot \prod_{t=1}^{T} p(x_t | z_t) \\
p(z_1) &= \text{Norm}(z_1 | 0, Q_0)
\end{align*}
\]

- HMM structure with Gaussian conditional distributions:

\[
p(z_{t+1} | z_t) = \text{Norm}(z_{t+1} | Az_t, Q) \quad p(x_t | z_t) = \text{Norm}(x_t | Wz_t, R)
\]

- Equivalent, system dynamics viewpoint:

\[
    z_{t+1} = Az_t + w_t, \quad w_t \sim \text{Norm}(0, Q) \\
x_t = Wz_t + v_t, \quad v_t \sim \text{Norm}(0, R)
\]

- Hidden states and observations are jointly Gaussian, so all marginals are Gaussian (parameterized by mean & covariance)

- Posterior distribution of state at any time, given observations at any subset of other times, is Gaussian
\[ x_{t+1} = x_t + w_t \]

\[
\begin{bmatrix}
  x_{t+1} \\
  \delta_{t+1}
\end{bmatrix} = \begin{bmatrix}
  1 & 1 \\
  0 & 1
\end{bmatrix} \begin{bmatrix}
  x_t \\
  \delta_t
\end{bmatrix} + w_t
\]
State Space Inference: Kalman Filter

Key Markov Identities: From generative structure of HMM,

\[ p(z_{t+1} | z_t, x_t, \ldots, x_1) = p(z_{t+1} | z_t) \]

Prediction Step: Given current knowledge, what is next state?

\[ p(z_{t+1} | x_t, \ldots, x_1) = \text{Norm}(z_{t+1} | \mu_{t+1|t}, P_{t+1|t}) \]

\[ p(z_{t+1} | x_t, x_{t-1}, \ldots, x_1) = \int p(z_{t+1} | z_t)p(z_t | x_t, x_{t-1}, \ldots, x_1) \, dz_t \]

\[ p(z_{t+1} | x_t, \ldots, x_1) = \int \text{Norm}(z_{t+1} | A z_t, Q)\text{Norm}(z_t | \mu_{t|t}, P_{t|t}) \, dz_t = \text{Norm}(z_{t+1} | \mu_{t+1|t}, P_{t+1|t}) \]

\[ \mu_{t+1|t} = A \mu_{t|t} \]

\[ P_{t+1|t} = AP_{t|t}A^T + Q \]
State Space Inference: Kalman Filter

Prediction Step: Given current knowledge, what is next state?

\[ p(z_{t+1} \mid x_t, x_{t-1}, \ldots, x_1) = \text{Norm}(z_{t+1} \mid \mu_{t+1|t}, P_{t+1|t}) \]

\[ \mu_{t+1|t} = A\mu_{t|t} \quad P_{t+1|t} = AP_{t|t}A^T + Q \]

Update Step: What does latest observation tell us about state?

\[ p(z_{t+1} \mid x_{t+1}, \ldots, x_1) \propto \text{Norm}(x_{t+1} \mid Wz_{t+1}, R)\text{Norm}(z_{t+1} \mid \mu_{t+1|t}, P_{t+1|t}) \]

\[ p(z_{t+1} \mid x_{t+1}, x_t, \ldots, x_1) \propto p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid x_t, \ldots, x_1) \]

Key Markov Identities: From generative structure of HMM,

\[ p(z_{t+1} \mid z_t, x_t, \ldots, x_1) = p(z_{t+1} \mid z_t) \quad p(x_{t+1} \mid z_{t+1}, x_t, \ldots, x_1) = p(x_{t+1} \mid z_{t+1}) \]
State Space Inference: Kalman Filter

**Prediction Step:** Given current knowledge, what is next state?

\[
p(z_{t+1} | x_t, x_{t-1}, \ldots, x_1) = \text{Norm}(z_{t+1} | \mu_{t+1|t}, P_{t+1|t})
\]

\[
\mu_{t+1|t} = A \mu_{t|t}
\]

\[
P_{t+1|t} = A P_{t|t} A^T + Q
\]

**Update Step:** What does latest observation tell us about state?

\[
p(z_{t+1} | x_{t+1}, \ldots, x_1) \propto \text{Norm}(x_{t+1} | Wz_{t+1}, R)\text{Norm}(z_{t+1} | \mu_{t+1|t}, P_{t+1|t})
\]

\[
\mu_{t+1|t+1} = \mu_{t+1|t} + K_{t+1}(x_{t+1} - W \mu_{t+1|t})
\]

\[
P_{t+1|t+1} = P_{t+1|t} - K_{t+1} W P_{t+1|t}
\]

\[
K_{t+1} = P_{t+1|t} W^T (WP_{t+1|t} W^T + R)^{-1}
\]
State Space Inference: Kalman Smoother

**Forward Recursion:** Distribution of State Given Past Data
\[
p(z_1 | x_1) \propto p(z_1)p(x_1 | z_1) \propto \text{Norm}(z_1 | \mu_{1|1}, P_{1|1})
\]
\[
\mu_{t+1|t+1} = A\mu_{t|t} + K_{t+1}(x_{t+1} - WA\mu_{t|t})
\]

Kalman gain \( K \), and covariances \( P \), depend on noise statistics.

**Backward Recursion:** Likelihood of Future Data Given State
\[
p(x_{t+1}, \ldots, x_T | z_t) \propto \text{Norm}(z_t | \rho_t, S_t)
\]
A "backwards" variant of Kalman filter can recursively compute these likelihood functions, which have a Gaussian shape.

**Marginal:** Posterior distribution of state given all data
\[
p(z_t | x_1, \ldots, x_T) \propto \text{Norm}(z_t | \mu_{t|t}, P_{t|t})p(x_{t+1}, \ldots, x_T | z_t)
\]
Constant Velocity Tracking

Kalman Filter

Kalman Smoother

(K. Murphy, 1998)