石油科学通报
Petroleum Science Bulletin

基于贝叶斯概率矩阵分解的地震数据重建算法

侯思安，张峰，李向阳*

中国石油大学(北京)油气资源与工程国家重点实验室，北京 102249
* 通信作者, xyl1962@hotmail.com

收稿日期: 2018−04−04
国家自然科学基金项目(41474096) 和国家科技重大专项(2017ZX05018005) 联合资助

摘要　低秩矩阵分解是一种机器学习算法，近年来该算法在地震数据重建问题中得到了广泛的关注，大量的学者针对模型构建和最优化求解等问题开展了研究。但是精确的求解低秩矩阵分解问题还需要知道规则化参数，而规则化参数又与地震数据体的均值和方差等统计学参数直接相关，又因为数据缺失和随机噪音等因素，这些参数是无法精确获取的。针对这一问题，本文引入了贝叶斯概率矩阵分解算法，通过对均值和方差进行随机模拟，并计算相应的概率密度函数，从而实现自适应的选取最优数据重建结果。合成地震记录和实际地震数据测试表明本文方法可以有效提高地震数据插值重建的精度和稳定性。

关键词　数据重建；机器学习；低秩矩阵分解；贝叶斯原理；马尔科夫蒙托卡罗方法

0　引言

地震数据重建是提高资料品质和降低勘探综合成本的关键技术。经典的数据重建算法主要有预测滤波算法(Prediction Filter)和促稀疏反演算法(Sparsity Promoting Inversion)：预测滤波算法是将含有缺失道的地震数据从时间—空间域变换到频率—空间域，然后根据数据在频率—空间域中表现出来的线性特征进行数据重建[1];而促稀疏反演算法是假设原始的地震信号在变换域中是稀疏的，当数据存在缺失时稀疏性会变弱，因此可以通过求解一个 l_1 模规则化的最优化问题来恢复信号的稀疏特征实现数据重建[2−3]。这些算法的一个共同点是都依赖于解析数学变换如快速傅里叶变换(Fast Fourier Transform)[4]，拉东变换(Radon Transform)[5]，曲波变换(Curvelet Transform)[6]和地震小波变换(Seislet Transform)[7]等。但是受限于计算机的性能和实际地震资料的复杂性，解析数学变换已经很难满足石油工业对数据处理精度的需求。因此学术界将研究的重心转向了机器学习(Machine Learning)算法，希望能通过机器学习达到更好的数据重建效果。

就地震数据重建这个问题而言，主要的机器学习算法有稀疏字典学习算法(Sparsity Dictionary Learning)和低秩矩阵分解算法(Low-rank Matrix Factorization)。与数学变换算法相似，稀疏字典学习算法是通过输入数据自适应的生成一组变换基函数，在该基函数下原始的输入数据具有稀疏特征。目前计算精度最高的稀疏字典学习算法是k-SVD算法[8−9]，在使用该算法进行地震数据插值时[10−11]：第一步，对原始数据进行Patch处理[12]并初始化一组基函数(在k-SVD算法中基函数被称为字典)；第二步，求解 l_1 模规则化问题计算稀疏系数；第三步，分别对基函数的每一列进行更新，更新每一列时要对残差矩阵进行一次SVD分解，并用分解的第一列左特征向量矩阵更新基函数，用右特征向量矩阵的第一行和第一个特征值的乘积更新系数矩阵的对应行；第四步，循环步骤二至步骤四若干次直至收敛为止。可以看出当基函数具有 k 列时，该

引用格式：侯思安, 张峰, 李向阳. 基于贝叶斯概率矩阵分解的地震数据重建算法. 石油科学通报, 2018, 02: 154−166
HOU Sian, ZHANG Feng, LI Xiangyang. Seismic data reconstruction via a Bayesian probabilistic matrix factorization algorithm. Petroleum Science Bulletin, 2018, 02: 154-166. doi: 10.3969/j.issn.2096-1693.2018.02.016

算法每一次迭代都要进行 k 次SVD分解，巨大的计算量限制了该算法在实际资料处理中的应用。数据驱动的紧致框架(Data-Driven Tight Frame，DDTF)[13]则是一种改进的稀疏字典学习算法，该算法对基函数施加了紧致性约束条件，在每次迭代时仅需要一次SVD分解，极大地提高了数据处理的效率[14−15]。低秩矩阵分解算法则假设实际的地震信号具有低秩特征，但是当数据中存在缺失或噪音时，信号的秩会提高，因此可以设计一个降秩的优化算法实现地震数据重建[16]，常用于地震数据处理的减秩算法有阻尼减秩算法(Damped Rank-reduction Method)[12]，特征值收缩算法(Singular Value Shrinkage)[17]和多通道奇异谱分析算法(Multichannel Singular Spectrum Analysis, MSSA)[18]等。此外实际用于地震数据插值的都是4D或5D数据体，针对这种高维信号可以采用构建Hankel矩阵对数据降维[17]或直接应用张量分解算法进行求解[19−20]。

相比较而言低秩矩阵分解算法在求解最优化问题时不用处理 l_1 模规则化，也不是必须要用SVD分解，因此在计算效率上具有一定的优势。但是在求解该算法时需要考虑多个最优化参数，多数情况下是通过设置不同的参数进行数据处理然后人工选取较优的结果，这样做会增大数据处理的计算量，且处理精度也难以保障。针对这一问题，Salakhutdinov和Mnih在求解推荐系统(Recommendation System)的矩阵分解问题时提出了贝叶斯概率矩阵分解(Bayesian Probabilistic Matrix Factorization，BPMF)算法[21]，该算法可以对基函数和系数的均值和方差等统计学参数进行随机模拟，通过计算不同参数的概率密度函数自适应的选取最优结果，Netflix问题测试表明该方法具有良好的应用效果。本文将该算法引入到地震数据重建问题中，大量的数值测试表明该算法可以提高数据处理的精度和稳定性。本文首先介绍了地震数据矩阵分解的概率解释和BPMF算法的原理；然后，通过单道子波、合成地震记录和实际资料对该方法进行测试；最后，对算法进行总结和展望。

1 方法原理

1.1 基于概率矩阵分解的地震数据插值算法

基于概率矩阵分解(Probabilistic Matrix Factorization，PMF)[22]的地震数据插值算法通常假设实际地震信号矩阵是低秩的，并且可以表示为两个矩阵的乘积形式：

$$\boldsymbol{X}=\boldsymbol{MA}+\boldsymbol{E} \tag{1}$$

其中，$\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$ 表示含有随机缺失的地震数据，n 和 m 分别表示矩阵的行数和列数；$\boldsymbol{M}\in\boldsymbol{R}^{n\times k}$ 和 $\boldsymbol{A}\in\boldsymbol{R}^{k\times m}$ 分别表示基函数和对应的系数，理论上 k 等于数据 $\boldsymbol{X}$ 的秩；$\boldsymbol{E}$ 表示随机噪音矩阵。

由于基函数 $\boldsymbol{M}$ 和系数 $\boldsymbol{A}$ 都是未知的，很自然的假设的这两个矩阵的元素都是符合Gaussian分布的：

$$\begin{aligned}P\left(\boldsymbol{M}\mid\sigma_{\mathrm{M}}^2\right)&=\mathcal{N}\left(\boldsymbol{M}\mid 0,\sigma_{\mathrm{M}}^2\right)\\&=\prod_{i=0}^{n-1}\prod_{j=0}^{k-1}\mathcal{N}\left(\mu_{i,j}\mid 0,\sigma_{\mathrm{M}}^2\right)\end{aligned} \tag{2}$$

$$\begin{aligned}P\left(\boldsymbol{A}\mid\sigma_{\mathrm{A}}^2\right)&=\mathcal{N}\left(\boldsymbol{A}\mid 0,\sigma_{\mathrm{A}}^2\right)\\&=\prod_{i=0}^{k-1}\prod_{j=0}^{m-1}\mathcal{N}\left(a_{i,j}\mid 0,\sigma_{\mathrm{A}}^2\right)\end{aligned} \tag{3}$$

其中，$\mathcal{N}\left(y\mid\varphi,\sigma^2\right)$ 表示 y 符合均值为 φ，方差为 σ^2 的Gaussian分布；$\mu_{i,j}$ 和 $a_{i,j}$ 分别表示 $\boldsymbol{M}$ 和 $\boldsymbol{A}$ 的元素；σ_{M}^2 和 σ_{A}^2 分别表示 $\boldsymbol{M}$ 和 $\boldsymbol{A}$ 的方差；$\prod$ 表示乘积运算符。

因为随机噪音 $\boldsymbol{E}$ 也是满足Gaussian分布的，所以：

$$\begin{aligned}P\left(\boldsymbol{E}\mid\sigma_{\mathrm{E}}^2\right)&=P\left(\boldsymbol{X}-\boldsymbol{MA}\mid\sigma_{\mathrm{E}}^2\right)\\&=\mathcal{N}\left(\boldsymbol{X}-\boldsymbol{MA}\mid 0,\sigma_{\mathrm{E}}^2\right)\end{aligned} \tag{4}$$

其中，σ_E^2 表示随机噪音 $\boldsymbol{E}$ 的方差。

根据概率密度函数的平移关系有：

$$\begin{aligned}P\left(\boldsymbol{E}\mid\sigma_{\mathrm{E}}^2\right)&=\mathcal{N}\left(\boldsymbol{X}\mid\boldsymbol{MA},\sigma_{\mathrm{E}}^2\right)\\&=\prod_{i=0}^{n-1}\prod_{j=0}^{m-1}\mathcal{N}\left(x_{i,j}\mid x_{i,j}^{*},\sigma_{\mathrm{E}}^2\right)^{I_{i,j}}\end{aligned} \tag{5}$$

其中，$x_{i,j}^{*}$ 表示重建地震数据 $\boldsymbol{X}^{*}=\boldsymbol{MA}$ 的元素；上标 $I_{i,j}$ 用于标记数据的缺失信息，当该元素缺失时 $I_{i,j}=0$，否则 $I_{i,j}=1$。

因此基于PMF的地震数据重建等价于在采集数据 X 已知时，求取 $\boldsymbol{M}$ 和 $\boldsymbol{A}$ 的最大后验概率：

$$P\left(\boldsymbol{M},\boldsymbol{A}\mid\boldsymbol{X}\right)=\frac{P\left(\boldsymbol{M},\boldsymbol{A},\boldsymbol{X}\right)}{P\left(\boldsymbol{X}\right)} \tag{6}$$

因为 $\boldsymbol{X}$ 是已知的，因此概率密度函数 $P\left(\boldsymbol{X}\right)$ 是一个常数，所以有：

$$\begin{aligned}&P\left(\boldsymbol{M},\boldsymbol{A}\mid\boldsymbol{X}\right)\propto P\left(\boldsymbol{M},\boldsymbol{A},\boldsymbol{X}\right)\\&=P\left(\boldsymbol{X}\mid\sigma_{\mathrm{E}}^2\right)P\left(\boldsymbol{M}\mid\sigma_{\mathrm{M}}^2\right)P\left(\boldsymbol{A}\mid\sigma_{\mathrm{A}}^2\right)\end{aligned} \tag{7}$$

对公式(7)求对数并带入概率密度函数(2)、(3)、(5)和高斯分布函数 $\mathcal{N}\left(y\mid\varphi,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(y-\varphi)^2}{-2\sigma^2}}$，可以得到：

$$\begin{aligned}\ln P(\boldsymbol{M},\boldsymbol{A}\mid\boldsymbol{X}) &= \sum_{i=0}^{n-1}\sum_{j=0}^{m-1} I_{i,j} ln\mathcal{N}\left(x_{ij}\mid x_{i,j}^{*},\sigma_{\mathrm{E}}^{2}\right)\\ &+ \sum_{i=0}^{n-1}\sum_{j=0}^{k-1} ln\mathcal{N}\left(\mu_{i,j}\mid 0,\sigma_{\mathrm{M}}^{2}\right)\\ &+ \sum_{i=0}^{k-1}\sum_{j=0}^{m-1} ln\mathcal{N}\left(a_{i,j}\mid 0,\sigma_{\mathrm{A}}^{2}\right)+c\end{aligned} \tag{8}$$

其中，c 是用于平衡方程(8)左右两端的一个常数。

根据公式(8)，在采集数据 X 已知时，求取基函数 $\boldsymbol{M}$ 和系数 $\boldsymbol{A}$ 的最大后验概率等价于求解如下的最优化问题：

$$\min_{\boldsymbol{M},\boldsymbol{A}}\frac{1}{2}\sum_{i,j\in\Omega}\left(x_{i,j}^{*}-x_{i,j}\right)^{2}+\lambda_{\mathrm{M}}\|\boldsymbol{M}\|_{2}^{2}+\lambda_{\mathrm{A}}\|\boldsymbol{A}\|_{2}^{2} \tag{9}$$

其中，$x_{i,j}$ 和 $x_{i,j}^{*}$ 分别表示采集得到的地震数据和重建的地震数据；$\lambda_{\mathrm{M}}=\dfrac{\sigma_{\mathrm{E}}^{2}}{\sigma_{\mathrm{M}}^{2}}$ 和 $\lambda_{\mathrm{A}}=\dfrac{\sigma_{\mathrm{E}}^{2}}{\sigma_{\mathrm{A}}^{2}}$ 表示最优化权系数，σ_{M}^{2}、σ_{A}^{2} 和 σ_{E}^{2} 表示基函数 $\boldsymbol{M}$、系数 $\boldsymbol{A}$ 和随机噪音 $\boldsymbol{E}$ 的方差；$\|\ \|_{2}$ 表示 l_{2} 模规则化；Ω 表示采集数据的观测系统。

公式(9)描述了基于PMF的地震数据插值算法，在应用该算法进行数据重建时需要提前知道规则化参数 λ_{M} 和 λ_{A}，而这两个参数又和基函数 $\boldsymbol{M}$、系数 $\boldsymbol{A}$ 和随机噪音 $\boldsymbol{E}$ 的方差 σ_{M}^{2}、σ_{A}^{2} 和 σ_{E}^{2} 相关。其中随机噪音方差 σ_{E}^{2} 可以通过分析地震数据进行估算，但是 σ_{A}^{2} 和 σ_{E}^{2} 通常是无法获取的。一种普遍的做法是尝试用不同的参数进行数据处理，然后选取其中效果最好的作为最后的处理结果，但是这样做的计算量是比较大的。因此，本文将贝叶斯概率矩阵分解算法引入到地震数据插值问题中，基于该算法可以有效地解决规则化参数不确定的问题。

1.2 基于贝叶斯概率矩阵分解的地震数据插值算法

贝叶斯概率矩阵分解(Bayesian Probabilistic Matrix Factorization，BPMF)的核心原理是假设决定基函数 $\boldsymbol{M}$ 的系数 $\boldsymbol{A}$ 分布特征的超参数 $\Theta_{\mathrm{M}}=\{\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\}$ 和 $\Theta_{\mathrm{A}}=\{\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\}$ 是符合Gaussian-Wishart分布的，通过马尔科夫蒙托卡罗(Markov Chain Monte Carlo，MCMC)方法对不同参数的结果进行随机模拟，并计算不同参数对应的重建结果的概率密度函数自适应的选取最优结果。BPMF的示意图如图 1 所示。

根据前文的阐述，在BPMF算法中基函数 $\boldsymbol{M}$ 和系数 $\boldsymbol{A}$ 满足如下的高斯分布：

$$P(\boldsymbol{M}\mid\xi_{\mathrm{M}},\Lambda_{\mathrm{M}})=\prod_{i=0}^{n-1}\mathcal{N}\left(\boldsymbol{M}_{i,:}\mid\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}^{-1}\right) \tag{10}$$

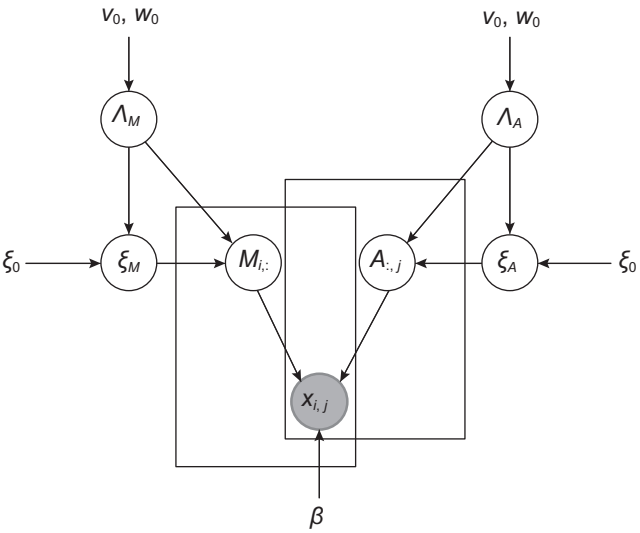

图 1 贝叶斯概率矩阵分解示意图

Fig. 1 The graphical model for Bayesian Probabilistic Matrix Factorization

$$P(\boldsymbol{A}\mid\xi_{\mathrm{A}},\Lambda_{\mathrm{A}})=\prod_{j=0}^{m-1}\mathcal{N}\left(\boldsymbol{A}_{:,j}\mid\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}^{-1}\right) \tag{11}$$

其中，$\boldsymbol{M}_{i,:}$ 表示基函数矩阵 $\boldsymbol{M}$ 的第 i 行，$\boldsymbol{A}_{:,j}$ 表示系数矩阵 $\boldsymbol{A}$ 的第 j 列，$\Theta_{\mathrm{M}}=\{\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\}$ 和 $\Theta_{\mathrm{A}}=\{\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\}$ 是决定基函数 $\boldsymbol{M}$ 和系数 $\boldsymbol{A}$ 分布特征的超参数，这两个超参数都符合Gaussian-Wishart分布：

$$\begin{aligned}P(\Theta_{\mathrm{M}}\mid\Theta_{0}) &= P(\xi_{\mathrm{M}}\mid\Lambda_{\mathrm{M}})P(\Lambda_{\mathrm{M}})\\ &= \mathcal{N}\left(\xi_{\mathrm{M}}\mid\xi_{0},(\beta_{0}\Lambda_{\mathrm{M}})^{-1}\right)\mathcal{W}(\Lambda_{\mathrm{M}}\mid\boldsymbol{W}_{0},v_{0})\end{aligned} \tag{12}$$

$$\begin{aligned}P(\Theta_{\mathrm{A}}\mid\Theta_{0}) &= P(\xi_{\mathrm{A}}\mid\Lambda_{\mathrm{A}})P(\Lambda_{\mathrm{A}})\\ &= \mathcal{N}\left(\xi_{\mathrm{A}}\mid\xi_{0},(\beta_{0}\Lambda_{\mathrm{A}})^{-1}\right)\mathcal{W}(\Lambda_{\mathrm{A}}\mid\boldsymbol{W}_{0},v_{0})\end{aligned} \tag{13}$$

其中，$\Theta_{0}=\{\xi_{0},v_{0},\boldsymbol{W}_{0}\}$，$\xi_{0}$ 等于 0，v_{0} 等于基函数的列数 k，$\boldsymbol{W}_{0}$ 为大小是 $k\times k$ 的单位矩阵；$\mathcal{W}(\Lambda\mid\boldsymbol{W}_{0},v_{0})$ 表示Wishart分布：

$$\mathcal{W}(\Lambda\mid\boldsymbol{W}_{0},v_{0})=\frac{1}{C}|\Lambda|^{(v_{0}-D-1)/2}e^{-\frac{1}{2}Tr\left(\boldsymbol{W}_{0}^{-1}\Lambda\right)} \tag{14}$$

其中，c 表示归一化参数，Tr 表示矩阵的迹。

在进行缺失地震数据重建时，可以利用贝叶斯边缘化处理(Marginalizing)来提高插值算法的精度。基本原理就是对任何可能出现的超参数 $\{\Theta_{\mathrm{M}},\Theta_{\mathrm{A}}\}$ 分别计算数据重建的结果，并对所有的结果进行加权求和，求和的权系数就是每个超参数对应的概率密度函数，因此基于BPMF的地震数据重建可以表示为如下的一个积分：

$$\begin{aligned}P(\boldsymbol{X}^{*}\mid\boldsymbol{X},\Theta_{0}) = \iint & P(\boldsymbol{X}^{*}\mid\boldsymbol{M},\boldsymbol{A})P(\boldsymbol{M},\boldsymbol{A}\mid\boldsymbol{X},\Theta_{\mathrm{M}},\Theta_{\mathrm{A}})\\ & P(\Theta_{\mathrm{M}},\Theta_{\mathrm{A}}\mid\Theta_{0})\mathrm{d}\{\boldsymbol{M},\boldsymbol{A}\}\mathrm{d}\{\Theta_{\mathrm{M}},\Theta_{\mathrm{A}}\}\end{aligned} \tag{15}$$

直接求解积分函数(15)几乎是不可能的，因此Salakhutdinov和Mnih提出了基于马尔科夫蒙托卡罗(Markov Chain Monte Carlo，MCMC)的求解方法。在MCMC的框架下，积分函数(15)可以近似如下的求和函数：

$$P\left(\boldsymbol{X}^{*}\mid\boldsymbol{X},\Theta_{0}\right)=\frac{1}{K}\sum_{k=0}^{K-1}P\left(x_{i,j}^{*}\mid\boldsymbol{M}^{(k)},\boldsymbol{A}^{(k)}\right)\tag{16}$$

其中，$\left\{\boldsymbol{M}^{(k)},\boldsymbol{A}^{(k)}\right\}$表示马尔科夫随机采样序列，序列的长度是$K$。

本文使用Gibbs方法对公式(16)进行随机采样。由于地震数据插值仅需要概率最大的基函数$\boldsymbol{M}$和系数$\boldsymbol{A}$，所以实际计算时不需要完整的求解概率密度函数(16)，而是通过多次迭代使得其达到稳定即可，那么基于BPMF的地震数据重建流程如下：

第一步，初始化基函数$\boldsymbol{M}^{(0)}$和系数$\boldsymbol{A}^{(0)}$，初始化时可以采用其他算法求得的结果或直接采用随机函数生成得到，对输入地震数据进行Patch处理[12]；

第二步，更新超参数$\Theta_{\mathrm{M}}=\{\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\}$和$\Theta_{\mathrm{A}}=\{\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\}$，由于基函数$\boldsymbol{M}$和系数$\boldsymbol{A}$在这一步中是已知的，超参数$\Theta_{\mathrm{M}}=\{\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\}$和$\Theta_{\mathrm{A}}=\{\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\}$的概率密度函数表示为：

$$\begin{aligned}&P\left(\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\mid M,\Theta_{0}\right)\\&\propto P\left(M\mid\xi_{\mathrm{M}},\Lambda_{\mathrm{M}},\Theta_{0}\right)P\left(\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\mid\Theta_{0}\right)\\&=\mathcal{N}\left(\xi_{\mathrm{M}}\mid\xi_{0}^{*},\left(\beta_{0}^{*}\Lambda_{\mathrm{M}}\right)^{-1}\right)\mathcal{W}\left(\Lambda_{\mathrm{M}}\mid\boldsymbol{W}_{0}^{*},v_{0}^{*}\right)\end{aligned}\tag{17}$$

其中，

$$\xi_{0}^{*}=\frac{\beta_{0}\xi_{0}+K\bar{\boldsymbol{M}}}{\beta_{0}+K}\quad\beta_{0}^{*}=\beta_{0}+K\quad v_{0}^{*}=v_{0}+K$$

$$\left(\boldsymbol{W}_{0}^{*}\right)^{-1}=\boldsymbol{W}_{0}^{-1}+K\bar{U}+\frac{\beta_{0}K}{\beta_{0}+K}\left(\xi_{0}-\bar{\boldsymbol{M}}\right)\left(\xi_{0}-\bar{\boldsymbol{M}}\right)^{T}$$

$$\bar{\boldsymbol{M}}=\frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{M}^{(k)}\quad\bar{U}=\frac{1}{K}\sum_{i=0}^{n-1}\left(\boldsymbol{M}^{(k)}\right)\left(\boldsymbol{M}^{(k)}\right)^{T}$$

同理：

$$\begin{aligned}&P\left(\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\mid\boldsymbol{A},\Theta_{0}\right)\\&\propto P\left(\boldsymbol{A}\mid\xi_{\mathrm{A}},\Lambda_{\mathrm{A}},\Theta_{0}\right)P\left(\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}\mid\Theta_{0}\right)\\&=\mathcal{N}\left(\xi_{\mathrm{A}}\mid\xi_{0}^{*},\left(\beta_{0}^{*}\Lambda_{\mathrm{A}}\right)^{-1}\right)\mathcal{W}\left(\Lambda_{\mathrm{A}}\mid\boldsymbol{W}_{0}^{*},v_{0}^{*}\right)\end{aligned}\tag{18}$$

其中，

$$\xi_{0}^{*}=\frac{\beta_{0}\xi_{0}+K\bar{\boldsymbol{A}}}{\beta_{0}+K}\quad\beta_{0}^{*}=\beta_{0}+K\quad v_{0}^{*}=v_{0}+K$$

$$\left(\boldsymbol{W}_{0}^{*}\right)^{-1}=\boldsymbol{W}_{0}^{-1}+K\bar{V}+\frac{\beta_{0}K}{\beta_{0}+K}\left(\xi_{0}-\bar{\boldsymbol{A}}\right)\left(\xi_{0}-\bar{\boldsymbol{A}}\right)^{T}$$

$$\bar{A}=\frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{A}^{(k)}\quad\bar{V}=\frac{1}{K}\sum_{j=0}^{m-1}\left(\boldsymbol{A}^{(k)}\right)^{T}\left(\boldsymbol{A}^{(k)}\right)$$

第三步，更新基函数$\boldsymbol{M}$，根据贝叶斯公式：

$$\begin{aligned}&P\left(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)P\left(\boldsymbol{X}\mid\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\\&=P\left(\boldsymbol{X}\mid\boldsymbol{M},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)P\left(\boldsymbol{M}\mid\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\end{aligned}\tag{19}$$

因为此时采集数据$\boldsymbol{X}$，系数矩阵$\boldsymbol{A}$和超参数$\Theta_{\mathrm{M}}=\{\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\}$都是已知的，所以条件概率密度函数$P\left(\boldsymbol{X}\mid\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)$是一个常数，那么：

$$\begin{aligned}&P\left(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\propto\\&P\left(\boldsymbol{X}\mid\boldsymbol{M},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\cdot\\&P\left(\boldsymbol{M}\mid\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\end{aligned}\tag{20}$$

又因为基函数矩阵$\boldsymbol{M}$和系数矩阵$\boldsymbol{A}$是无关的：

$$\begin{aligned}&P\left(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\propto\\&P\left(\boldsymbol{X}\mid\boldsymbol{M},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\cdot\\&P\left(\boldsymbol{M}\mid\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\end{aligned}\tag{21}$$

所以：

$$\begin{aligned}&P\left(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{A},\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}\right)\\&\propto\prod_{i=0}^{n-1}\prod_{j=0}^{m-1}\mathcal{N}\left(x_{i,j}\mid\boldsymbol{M},\boldsymbol{A},\sigma^{2}\right)^{I_{i,j}}\\&\prod_{i=0}^{n-1}\mathcal{N}\left(\boldsymbol{M}_{i,:}\mid\xi_{\mathrm{M}},\Lambda_{\mathrm{M}}^{-1}\right)\end{aligned}\tag{22}$$

第四步，更新系数$\boldsymbol{A}$，与第三步同理可以得：

$$\begin{aligned}&P\left(\boldsymbol{X}\mid\boldsymbol{M},\boldsymbol{A},\xi_{\mathrm{X}},\Lambda_{\mathrm{X}}\right)\\&\propto\prod_{i=0}^{n-1}\prod_{j=0}^{m-1}\mathcal{N}\left(x_{i,j}\mid\boldsymbol{M},\boldsymbol{A},\sigma^{2}\right)^{I_{i,j}}\\&\prod_{j=0}^{m-1}\mathcal{N}\left(\boldsymbol{A}_{:,j}\mid\xi_{\mathrm{A}},\Lambda_{\mathrm{A}}^{-1}\right)\end{aligned}\tag{23}$$

最后，重复步骤二至步骤四直至基函数$\boldsymbol{M}$和系数$\boldsymbol{A}$达到稳定为止。

数据重建的流程如图 2 所示。

1.3 贝叶斯概率矩阵分解算法的稳定性分析

BPMF算法是一种随机缺失数据的恢复算法，要对缺失数据进行精确的恢复除了要求原始数据是低秩的，还要满足一定的采样条件。Candes和Recht对这一问题进行了深入研究，当随机采样满足如下的条件时即可用低秩算法进行数据重建[16]，具体条件如下：

$$e\geqslant Cf^{6/5}r\log f\tag{24}$$

其中，e表示随机采样的数据点数；$f=\max\left(m,n\right)$表示原始数据矩阵行数或列数的大值；r表示原始数据矩阵的秩；C表示一个数值常数。根据采样条件，当

原始数据的秩 r 比较小时($Cf^{6/5}r\log f<m\times n$)，可以用随机采样的数据精确地恢复出原始数据；但是当 r 比较大时($Cf^{6/5}r\log f>m\times n$)，即便知道了所有有效信息，低秩算法也会造成原始数据的缺失。

图 3 表示用随机生成的低秩矩阵进行的BPMF方法稳定性测试。在测试中，原始数据矩阵是一个100×100 的方阵，矩阵的秩从 1 逐渐增加到 41，采样率从 0%逐渐增加到 100%，数据重建的效果使用峰值信噪比(公式 25)进行评估，并假设信噪比大于 15 时为重建成功，否则为失败。结果表明BPMF算法是一种精确稳定的低秩恢复算法，仅需要不到 20%的随机采样数据就可以对低秩数据进行恢复。

基于BPMF的地震数据插值算法，在理论上可以有效的减少数据处理过程中对优化参数的依赖，提高数值计算的稳定性和精度，并用一个随机数矩阵验证了方法的稳定性。但是真正的地震数据并不是完全随机缺失的(表现为在时间方向连续缺失，在空间方向随机缺失)，并且地震数据也不是完全低秩数据，这些问题可能会影响基于BPMF算法的地震数据重建效果。本文将通过合成地震数据和实际资料对本方法进行测试。

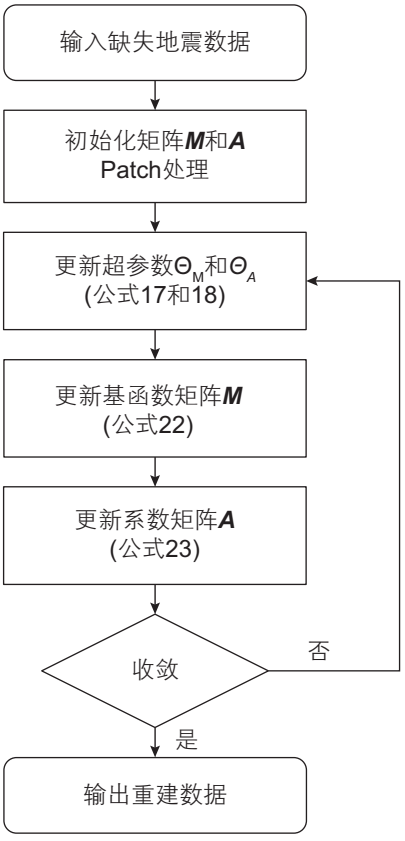

图 2 贝叶斯概率矩阵分解算法计算流程

Fig. 2 Workflow of Bayesian probabilistic matrix factorization algorithm

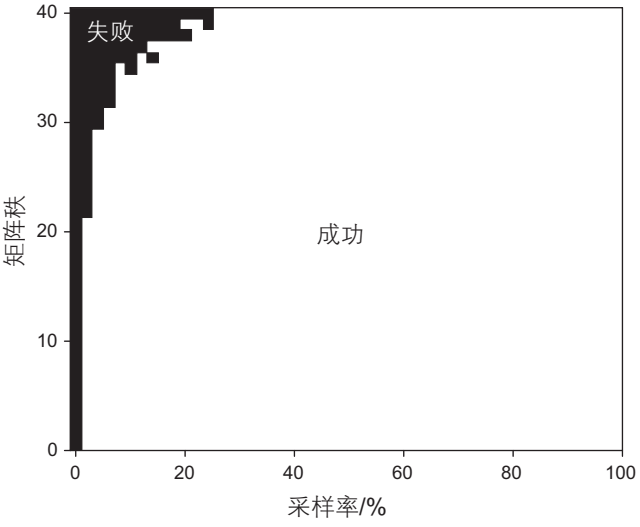

图 3 随机数矩阵重建测试：白色表示重建成功；黑色表示重建失败

Fig. 3 Illustration of random matrix recovery: white and black represents the successful recovery and failed recovery, respectively

2 方法测试

2.1 单道地震数据测试

首先通过随机缺失的单道地震子波数据对方法进行测试。单道记录如图 4a所示，道集总共有 401 个采样点，采样间隔为 1 ms，4 个独立的子波主频分别为 15 Hz、30 Hz、15 Hz和 30 Hz，振幅分别为 1.0、1.0、3.0 和 3.0，数据总计有 201 道随机缺失。数据重构结果如图 4b所示，4 条不同的线段分别表示Patch大小为 4、8、12 和 16 的重构子波(Patch处理类似于地震数据的时窗概念，是一种常用的图像处理手段，具体定义可参考文献 12)，图 4c和图 4d分别表示了0~200 ms和 200~400 ms的局部放大图。根据Patch=4和Patch=8 的重建结果，地震数据重建的Patch尺寸不能太小，否则会出现局部数据的采样不足，无法实现数据重建；同时对比Patch=12 和Patch=16 的重建结果，当Patch过大时，重建结果容易过度平滑；并且对比不同的子波结果发现，本文方法更容易使主频较高、振幅较强的数据产生平滑，因此在实际数据处理时需要尝试不同的Patch参数然后选取最优的计算结果。

2.2 合成地震记录测试

本文通过合成地震记录对提出方法进行测试。测试数据如图 5a所示，原始数据具有 201 个地震道，每

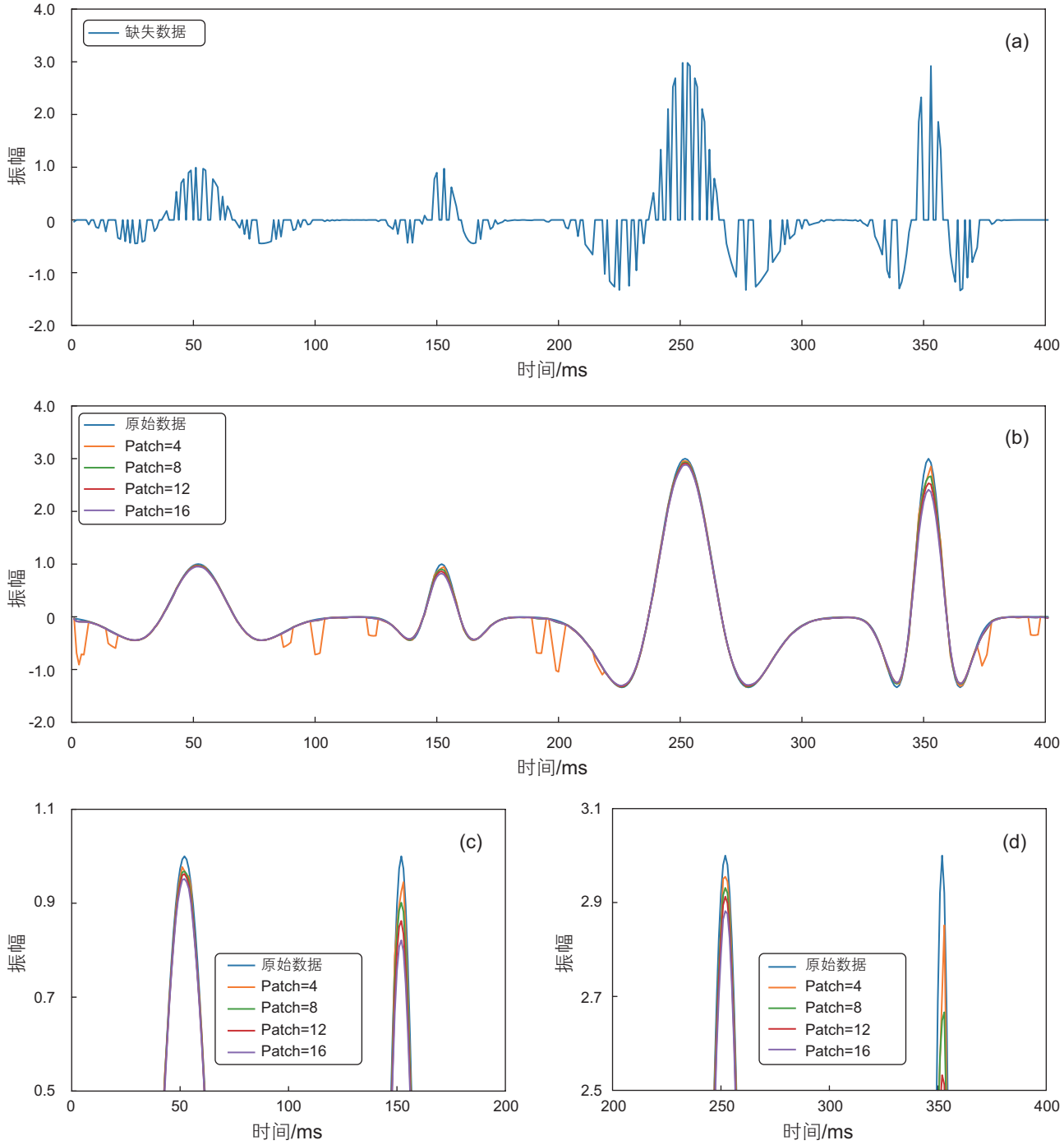

图 4　单道地震记录测试：(a)含有随机缺失的单道数据，从左到右的子波依次为：主频 15 Hz振幅 1.0，主频 30 Hz振幅 1.0、主频 15 Hz振幅 3.0 和主频 30 Hz振幅 3.0；(b)重建的单道数据；(c)0~200 ms局部放大图；(d)200~400 ms局部放大图

Fig. 4　Single trace test：(a)Random missing single trace data, wavelets parameters(from left to right): 15 Hz amp=1.0, 30 Hz amp=1.0, 15 Hz amp=3.0 and 30 Hz amp=3.0; (b) Reconstructed single trace data; (c) Zoom of reconstructed data(0~200ms); (d) Zoom of reconstructed data(200~400 ms)

个地震道具有 201 个采样点，空间和时间采样步长分别为 10 ms和 7 ms，随机选取其中的 40 道为缺失数据，如图 5b所示。在数据重建时选取基函数 $\boldsymbol{M}$ 的列数 k 等于 20，规则化参数 $\lambda_M = \lambda_A = 0.01$ 。图 5c、图 5d和图 5e分别表示基于Curvelet方法[23]，PMF方法和BPMF方法的重建结果；图 5f、图 5g和图 5h分别表示对应的重建数据残差。通过这一组数据对比表明，基于矩阵分解原理的PMF算法和BPMF算法比经典的Curvelet算法能更加有效地恢复地震数据，重建结果的残差明显减少。但是当数据在一定的区域内存在大

量的缺失时，PMF算法的重建效果出现了较为严重的下降，如图 5d的右侧所示。相比较而言，在相同的区域BPMF算法能较好的提高数据重建的效果，差剖面中没有明显的数据畸变，如图 5h所示。

图 6 展示了基于BPMF算法进行地震数据矩阵分解得到的基函数 $\boldsymbol{M}$，其中图 6a表示BPMF的基函数，图 6b表示离散余弦变换(Discrete Cosine Transform，DCT)的基函数，每一个单独的数据块代表基函数的一列。可以看出BPMF算法的基函数的每一个数据块都非常类似于地震数据的同相轴，区别在于振幅和相位的不同。而DCT基函数是具有解析数学表达式的传统数学变换，每个数据块都是固定的，与输入数据无关。

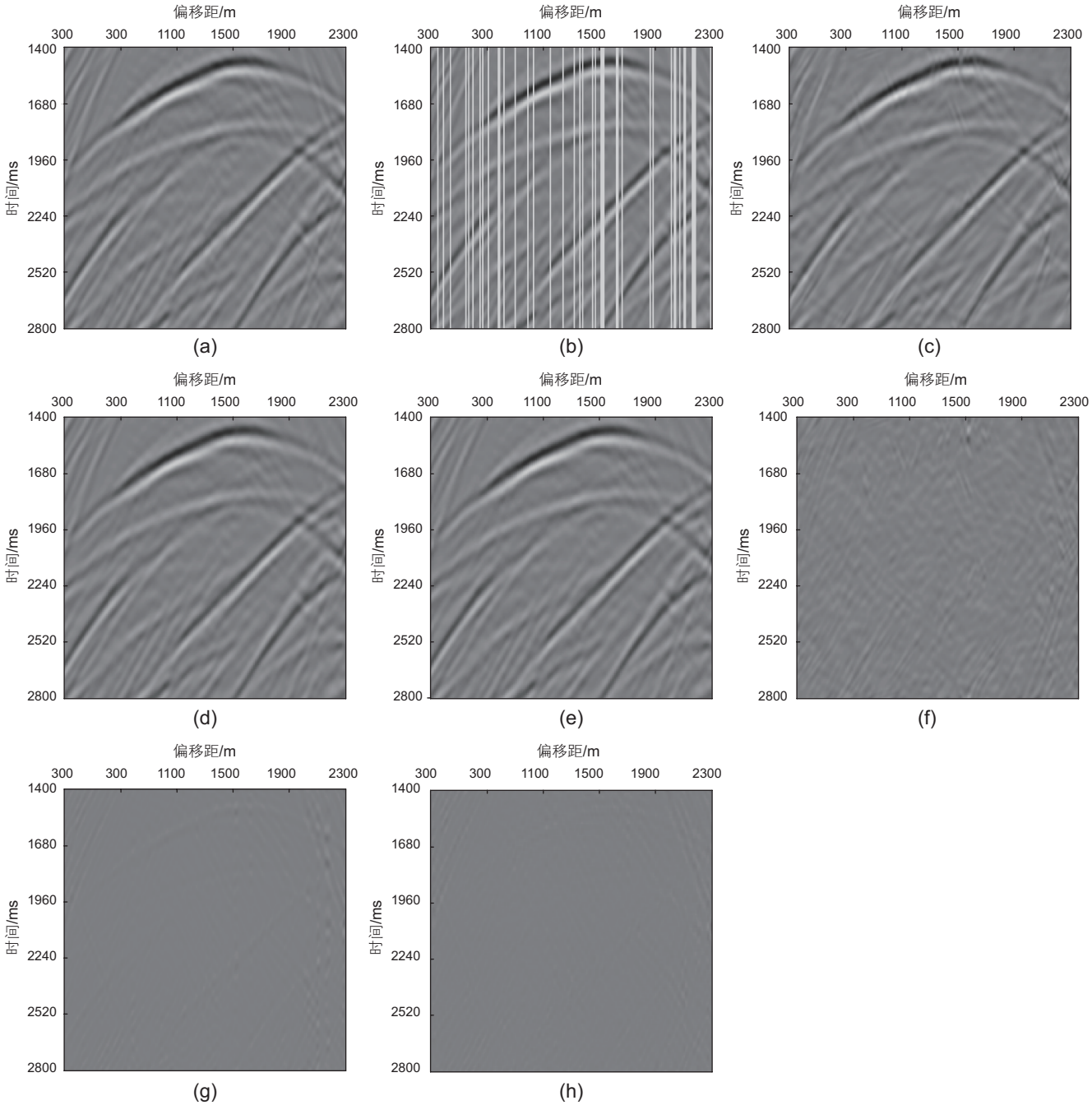

图 5 合成地震记录测试：(a)原始数据；(b)含随机缺失数据；(c)Curvelet重建数据；(d)PMF重建数据；(e)BPMF重建数据；(f)Curvelet重建差剖面；(g)PMF重建差剖面；(h)BPMF重建差剖面

Fig. 5 Synthetic data reconstruction test: (a) Original data; (d) Random missing data; (c) Reconstruction data via Curvelet; (d) Reconstruction data via PMF; (e) Reconstruction data via BPMF; (f) Residual of Curvelet reconstruction data; (g) Residual of PMF reconstruction data; (h) Residual of BPMF reconstruction data

通过对比可以看出矩阵分解算法可以根据输入数据自适应的构建基函数 $\boldsymbol{M}$，具有特征提取的功能。

为了更进一步的对比PMF算法和BPMF算法，测试了随机缺失道数从 10 道递增到 100 道，基函数 $\boldsymbol{M}$ 的列数 k 分别为 15、20 和 25，以及规则化参数 λ_M 和 λ_A 分别为 0.01、0.02 和 0.03 的情况，测试结果如图 7 所示，峰值信噪比的计算公式为：

$$PSNR = 10\log_{10}\frac{\left\|data_{orignal}\right\|^2}{\left\|data_{orignal} - data_{noisy}\right\|^2} \tag{25}$$

由于地震数据的复杂性以及信号噪音和缺失的干扰，这些重建参数是难以直接获取的，也没有一个明确的选取准则。但是通过这一组数据对比，当 k 从 15

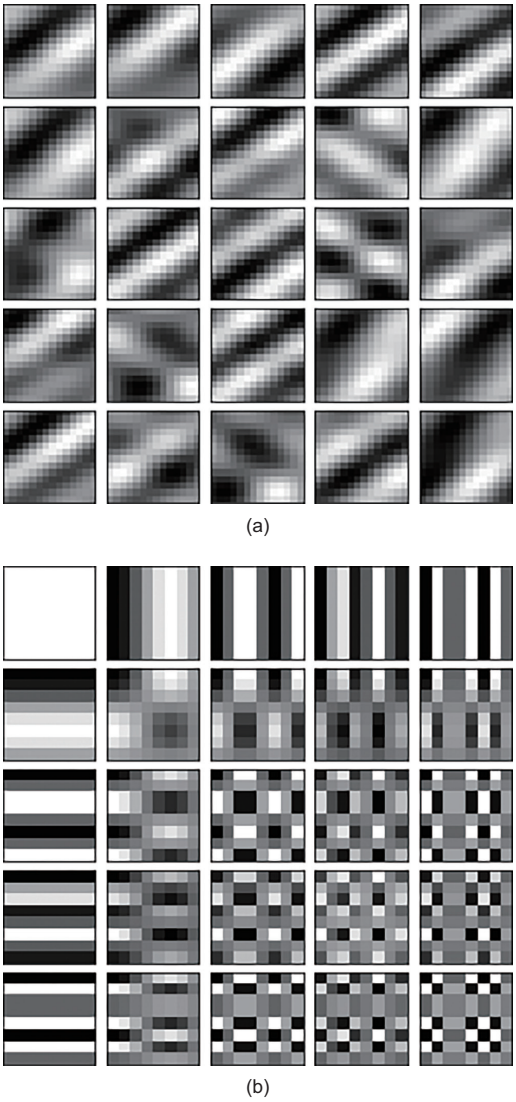

图 6 基函数示意图：(a)BPMF基函数示意图；(b)DCT基函数示意图

Fig. 6 Illustration of basis matrix: (a) Basis matrix of BPMF; (b) Basis matrix of DCT

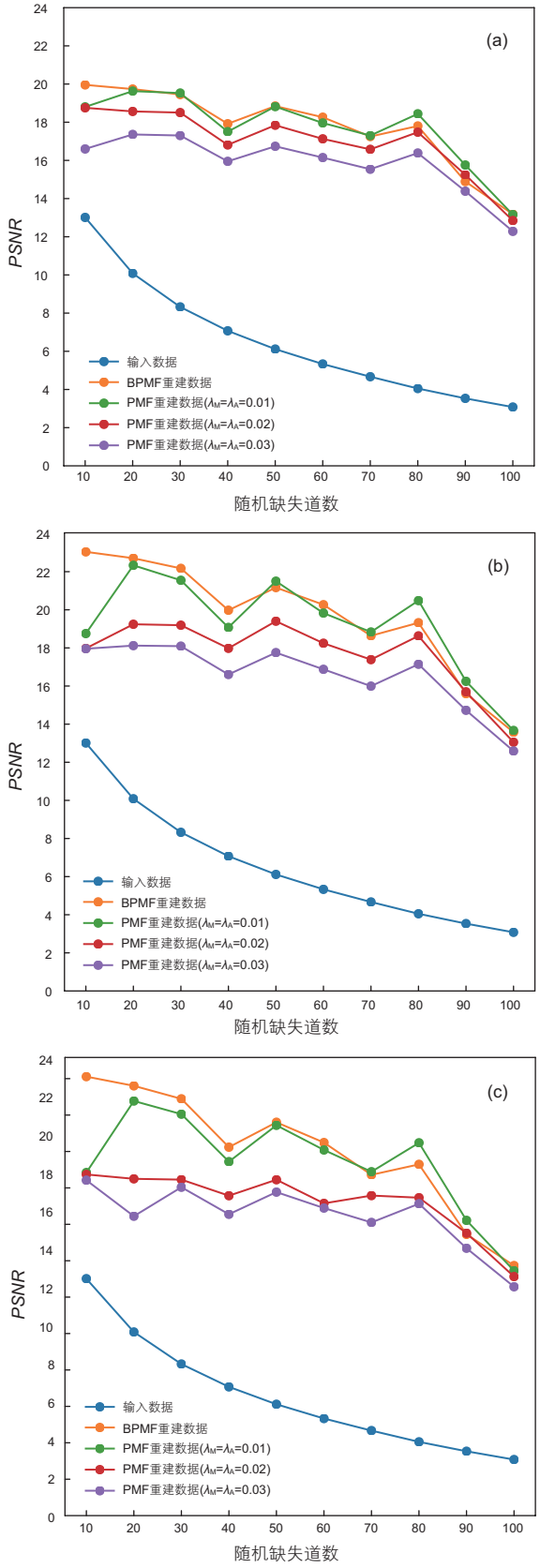

图 7 PMF和BPMF对比测试：(a) $k=15$ 测试结果；(b) $k=20$ 测试结果；(c) $k=25$ 测试结果

Fig. 7 Comparison PMF with BPMF: (a) Test result of $k=15$; (a) Test result of $k=20$; (a) Test result of $k=25$

逐渐增加到 25 并且规则化参数 λ_M 和 λ_A 从 0.03 逐渐减少到 0.01 时，PMF 算法数据重建的精度在不断提高，但是也出现了较大幅度的波动，这说明算法的稳定性也有所降低，但是 BPMF 算法始终保持了一个较高的

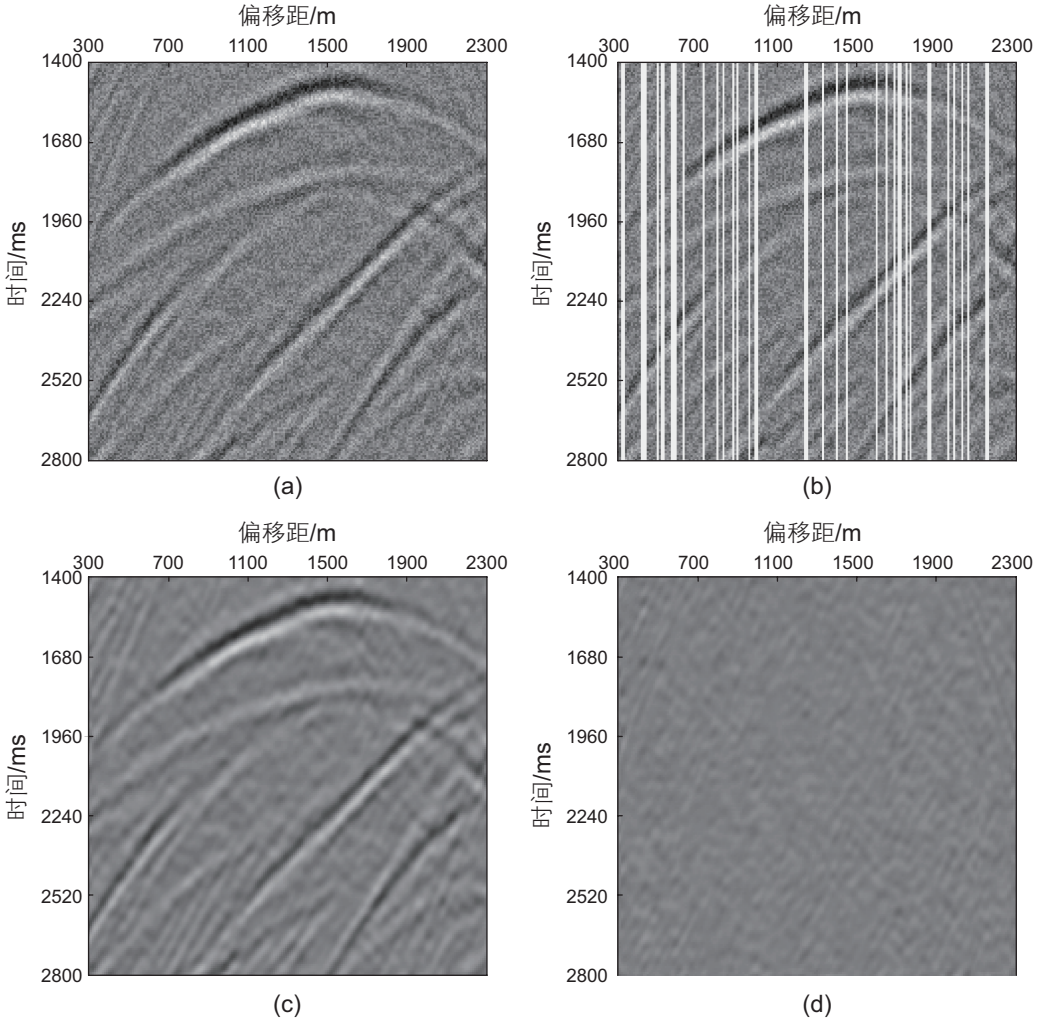

图 8 含噪音合成地震记录测试：(a) 原始数据；(b) 含随机缺失数据；(c)BPMF 重建数据；(d)BPMF 重建差剖面

Fig. 8 Synthetic noisy data reconstruction test: (a) Original data; (b) Random missing data; (c) Reconstruction data via BPMF; (d) Residual of BPMF reconstruction data

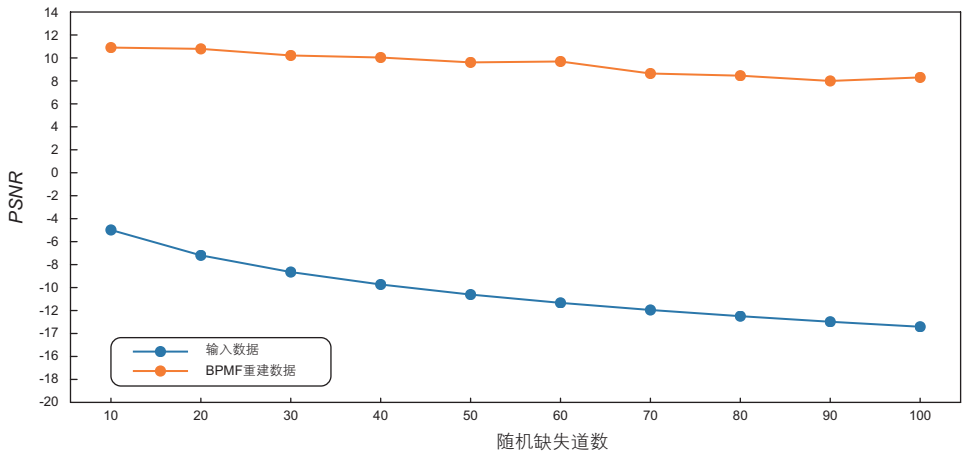

图 9 含噪音合成地震记录全部测试

Fig. 9 Overview of synthetic noisy data reconstruction test

计算精度和稳定性。

图 8 和图 9 则测试了在含有随机噪音的情况下，BPMF 算法的重建效果，其中图 8 展示了随机 25 道缺失的重建效果，图 9 展示了随机缺失道数从 10 道递增到 100 道时的全部重建结果。通过这一组测试可以看出当数据中存在随机噪音时，BPMF 算法依然可以稳定高效的对缺失地震数据进行重建。

2.3 实际地震记录测试

最后通过实际资料测试BPMF地震数据重建算法。测试使用的是海上OBC数据的一个剖面，该剖面有 101 个地震道，每个地震道具有 2501 个采样点，时间采样间隔为 2 ms，截取其中 1000 ms至 3000 ms的一段数据用于测试，如图 10a所示；测试时随机选取了其中的 38 道数据为缺失道，如图 10b所示；基于BPMF算法的重建结果($k = 25$)和差剖面分别如图 10c和图 10d所示；图 11 则分别展示了第 22 道、第 74 道和第 97 道重建数据和原始数据的单道记录。通过对比重建的剖面和单道记录，大部分缺失的地震数据被精确的恢复出来了，特别是对于主频较低的深层信号，

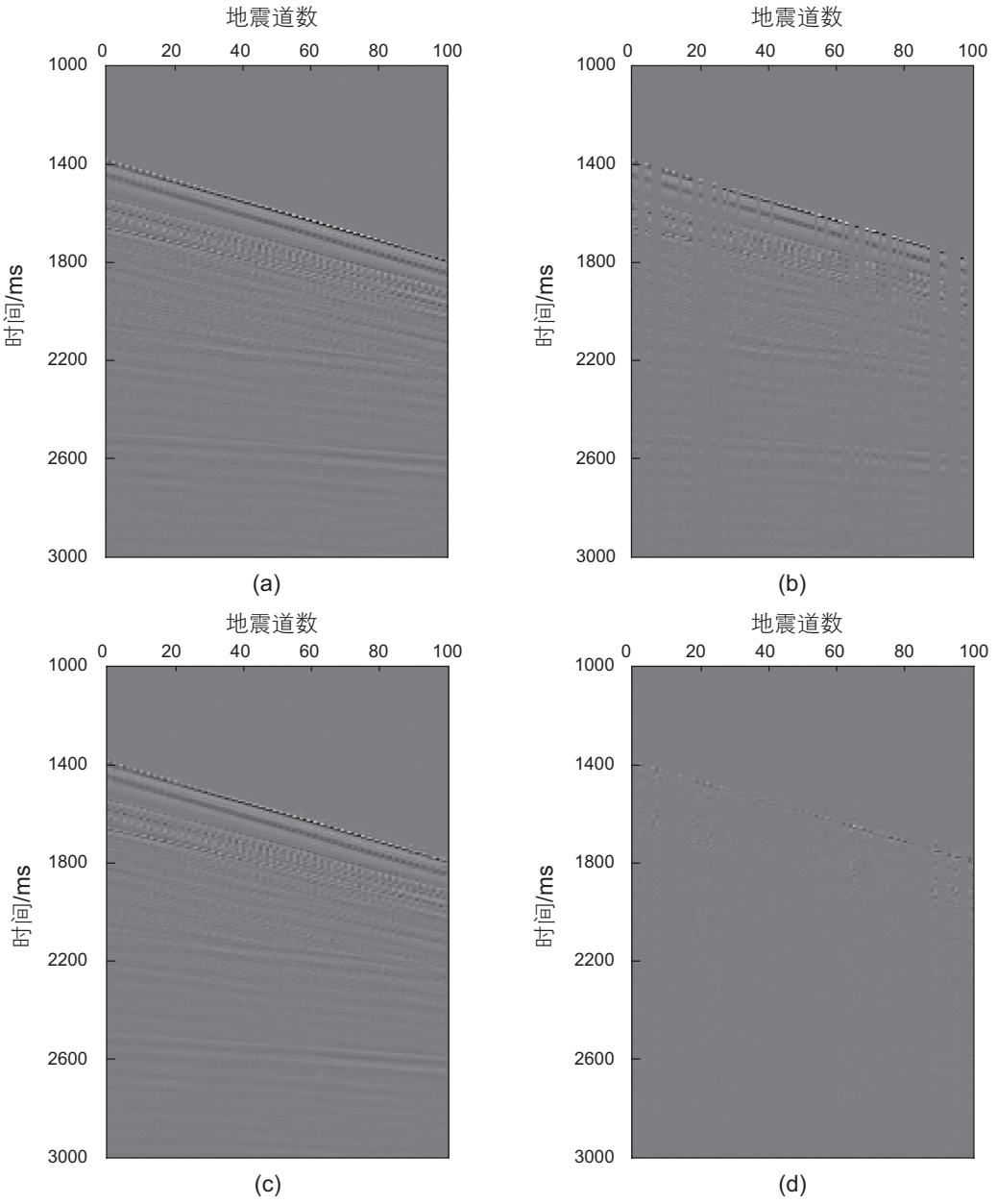

图 10 实际地震数据测试：(a)原始地震数据；(b)含缺失道地震数据；(c)BPMF重建地震数据；(d)BPMF 重建差剖面

Fig. 10 Real seismic data test: (a) Original seismic data; (b) Missing seismic data; (c) Reconstructed seismic data via BPMF; (d) Residual of BPMF reconstruction data

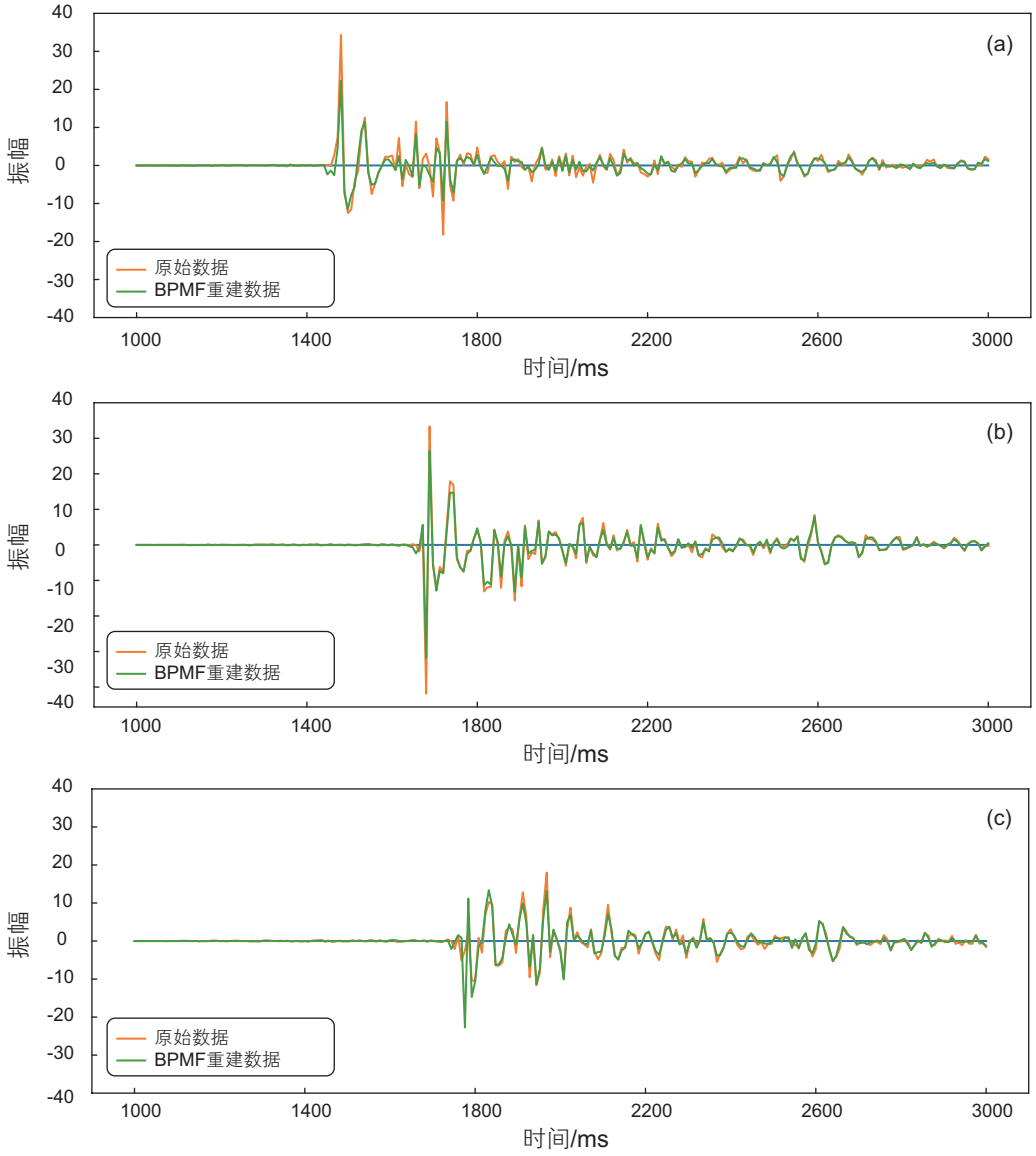

图 11 实际地震数据测试单道记录：(a) 第 22 道记录；(b) 第 74 道记录；(c) 第 97 道记录

Fig. 11 Trace data of real seismic data test: (a) No. 22 trace data; (b) No. 74 trace data; (c) No. 97 trace data

差剖面中几乎不含残留的有效信号，这说明本文提出的BPMF算法可以有效的对缺失地震数据进行重建。但是对能量较强的信号，数据恢复的结果中出现了比较明显的平滑，恢复信号的振幅能量降低了，这与单道记录测试结果是一致的，也是今后研究需要解决的问题。图 12 则展示了基于BPMF算法计算的基函数，该基函数依然体现出了非常明显的地震数据同相轴特征。

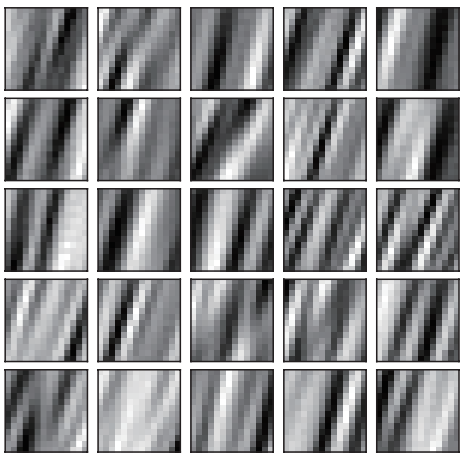

图 12 实际地震数据基函数

Fig. 12 Basis matrix of real seismic data

3 结论与展望

矩阵分解算法是在地震数据信号处理中非常具有研究价值的工业应用前景的机器学习算法，在应用该

算法时需要求解一个较为复杂的最优化问题，影响最终处理结果的有基函数 $\boldsymbol{M}$ 、系数 $\boldsymbol{A}$ 和随机噪音 $\boldsymbol{E}$ 的方差 σ_M^2 、 σ_A^2 和 σ_E^2 以及基函数的列数 k (理论上 k 应该等于实际地震数的秩) 等参数。本文针对 σ_M^2 和 σ_A^2 的选取问题开展了研究，通过引入马尔科夫蒙托卡罗方法对不同的参数进行随机模拟并计算相应的概率密度函数自适应的选取最优结果，合成地震数据和实际资料测试表明该算法在计算精度和稳定性方面均有所提高。

但是该算法也存在一些需要研究和解决的问题。首先，当数据中能量较强的同相轴，本文算法会对结果产生一定的平滑作用，导致重建数据失真，初步分析认为这与Patch处理有直接关系，但是如何解决这个问题还需要深入研究；然后贝叶斯算法可以根据输入数据自适应的生成一个基函数，该基函数表现出了非常明显的地震数据特征，是否可以利用该基函数以及如何利用该基函数还是一个需要深入研究的问题。

参考文献

[1] SPITZ S. Seismic trace interpolation in the F-X domain[J]. Geophysics, 1991, 56(6): 785−794.

[2] BAI L S, LIU Y K, LU H Y, et al. Curvelet-domain joint iterative seismic data reconstruction based on compressed sensing[J]. Chinese Journal of Geophysics, 2014, 57(9): 2937−2945.

[3] HERRMANN F J, HENNENFENT G. Non-parametric seismic data recovery with curvelet frames[J]. Geophysical Journal of the Royal Astronomical Society, 2010, 173(1): 233−248.

[4] ABMA R, KABIR N. 3D interpolation of irregular data with a POCS algorithm[J]. Geophysics, 2006, 71(6): E91.

[5] WANG J, NG M, PERZ M. Seismic data interpolation by greedy local Radon transform[J]. Geophysics, 2010, 75(6): WB225−WB234.

[6] CANDÈS E, DEMANET L, DONOHO D, et al. Fast discrete curvelet transforms[J]. multiscale modeling & simulation, 2006, 5(3): 861−899.

[7] FOMEL S, LIU Y. Seislet transform and Seislet frame[J]. Geophysics, 2010, 75(3): V25−V38.

[8] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311−4322.

[9] ELAD M, AHARON M. Image denoising via sparse and redundant representations over learned dictionaries[J]. IEEE Transactions on Image Processing, 2006, 15(12): 3736−3745.

[10] BECKOUCHE S, MA J. Simultaneous dictionary learning and denoising for seismic data[J]. Geophysics, 2014, 79(3): A27−A31.

[11] CHEN Y. Fast dictionary learning for noise attenuation of multidimensional seismic data[J]. Geophysical Journal International, 2017, 209(1): 21−31.

[12] MA J. Three-dimensional irregular seismic data reconstruction via low-rank matrix completion[J]. Geophysics, 2013, 78(5): V181−V192.

[13] CAI J F, JI H, SHEN Z, et al. Data-driven tight frame construction and image denoising[J]. Applied & Computational Harmonic Analysis, 2014, 37(1): 89−105.

[14] LIANG J, MA J, ZHANG X. Seismic data restoration via data-driven tight frame[J]. Geophysics, 2014, 79(3): V65−V74.

[15] YU S, MA J, ZHANG X, et al. Interpolation and denoising of high-dimensional seismic data by learning a tight frame[J]. Geophysics, 2015, 80(5): V119−V132.

[16] CANDÈS E J, RECHT B. Exact matrix completion via convex optimization[J]. Foundations of Computational Mathematics, 2008, 9(6): 717.

[17] CHEN K, SACCHI M D. Robust reduced-rank filtering for erratic seismic noise attenuation[J]. Geophysics, 2015, 80(1): V1−V11.

[18] CHEN Y, ZHANG D, JIN Z, et al. Simultaneous denoising and reconstruction of 5-D seismic data via damped rank-reduction method[J]. Geophysical Journal International, 2016, 206(3): 1695–1717.

[19] KREIMER N, SACCHI M D. A tensor higher-order singular value decomposition for prestack seismic data noise reduction and interpolation[J]. Geophysics, 2012, 77(3): V113−V122.

[20] GAO J, STANTON A, SACCHI M D. Parallel matrix factorization algorithm and its application to 5D seismic reconstruction and denoising[J]. Geophysics, 2015, 80(6): V173−V187.

[21] SALAKHUTDINOV R, MNIH A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo[C]. International Conference on Machine Learning, Helsinki, Fabianinkatu, ACM, 2008: 880−887.

[22] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization[C]. International Conference on Neural Information Processing Systems. Vancouver, Canada, 2007: 1257−1264.

[23] DAUBECHIES I, DEFRISE M, DE MOL C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint[J]. Communications on Pure & Applied Mathematics, 2004, 57(11): 1413−1457.

Seismic data reconstruction via a Bayesian probabilistic matrix factorization algorithm

HOU Sian, ZHANG Feng, LI Xiangyang

State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum-Beijing, Beijing 102249, China

Abstract Low-rank matrix factorization is a kind of machine learning algorithm. In recent years, the algorithm has received extensive attention in the problem of seismic data reconstruction. Much research related to model building and numerical calculations has been published. However, the exact solution of low-rank matrix factorization requires the regularization parameters, and the regularization parameters are directly related to the statistical parameters such as the mean and variance of the decomposed seismic data. But these parameters cannot be obtained precisely because of missing data and random noise. In order to solve this problem, this paper introduces the Bayesian probabilistic matrix factorization algorithm, which simulates the mean and variance randomly and calculates the optimal reconstruction result by calculating the probability density function. Synthetic seismic data and real seismic data tests indicate that the proposed method could improve the accuracy and stability of seismic data reconstruction.

Keywords data reconstruction; machine learning; low-rank matrix factorization; Bayes's theorem; Markov chain Monte Carlo

doi: 10.3969/j.issn.2096-1693.2018.02.016

(编辑 付娟娟)